Collaborative Learning: Aligning Goals and Outcomes

Mariel Werner

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2024-143

http://www2.eecs.berkeley.edu/Pubs/TechRpts/2024/EECS-2024-143.html

June 26, 2024

Collaborative Learning: Aligning Goals and Outcomes

by

Mariel Anne Farrar Werner

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Chair
Assistant Professor Jiantao Jiao
Assistant Professor Jacob Noah Steinhardt
Associate Professor Martin Jaggi

Summer 2024

Collaborative Learning: Aligning Goals and Outcomes

Abstract

Collaborative Learning: Aligning Goals and Outcomes

by

Mariel Anne Farrar Werner

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Michael I. Jordan, Chair

When multiple clients are collaboratively learning and training a shared model, incentives problems can arise. The clients may have different learning objectives and application domains, or they may be competitors whose participation in the learning system could reduce their competitive advantage. While collaborative learning is a powerful framework that leverages vast networks of compute and data to generate a better model for all, participants may defect from collaboration if their incentives are misaligned with the guarantees of the system. In this dissertation, I examine three areas where accounting for incentives is critical in designing an effective collaborative learning system. **I.** When clients in the system have heterogeneous data distributions and divergent learning tasks, full collaboration within the system can result in a global model which performs poorly for individual clients. Personalization of the global model to clusters of clients with similar learning objectives is a solution to this problem. We propose a personalization method which has optimal convergence guarantees and is provably robust to malicious attackers. **II.** Clients who are competitors may not want to participate in collaborative learning system if their contributions will benefit their competitors and disadvantage themselves. We design a collaborative learning scheme which guarantees that clients lose no utility by participating. Additionally, we show that even as clients focus on increasing their own revenues, their model qualities converge to the Nash bargaining solution, thus optimizing for joint surplus. **III.** Finally, privacy concerns are a major deterrent for joining collaborative learning systems. In the final chapter, we look at privacy dynamics in systems of learning agents more broadly. Specifically, we study a repeated-interaction game between potentially antagonistic learning agents – a buyer and a price-discriminating seller – and show that privacy-protecting behavior endogenously arises at equilibrium.

For Mommie, Oppie, Hélène, Lucien, and Andrée,
to whom I owe everything.

Contents

Acknowledgments

All my gratitude to my advisor Mike for his unwavering support. His integrity as an intellectual and as a person has been an inspiration since day one.

Thank you to Praneeth for opening my eyes to much of what I've written in this dissertation, and for showing me how to think deeply and patiently about hard problems.

Thank you to my thesis committee, Jiantao, Jacob, and Martin, for their wisdom and energy, and the paths they pave for their students.

Thank you to Martin (Wainwright) and Jiantao for conversations early on in my PhD that clarified my priorities.

Thank you to my cohort Anastasios, Reese, Mihaela, and Frances for solidarity in the early years of classes and exams, and to Anastasios, Reese, and Neha for many wonderful years of self-guided learning.

Thank you to Lie, Nivasini, Tiffany, Stephen, and Anastasios for all they taught me during our collaborations.

Thank you to my lab SAIL for its troves of intellectual power. Everything I learned in our meetings constituted a second PhD for me.

Thank you to Jean and Naomi for rounding out the sharp edges of doctoral life – to Jean for helping me navigate all the details and always being a reassuring presence, and to Naomi for making our lab such a welcoming and cheerful place.

I am with you, Nana. You supported my education and music every step of the way.

Mommie, Oppie, Hélène, Lucien, and Andrée, it is my greatest privilege in life to be your daughter and sister.

Mommie, you made it all possible.

Chapter 1

Introduction

Collaborative learning is a powerful distributed optimization framework that allows multiple clients in a network to train a single global model. It has many desirable properties. 1) Since clients do not transmit their raw data across the network, only a version of the global model which they have updated on their data, it guarantees a baseline of privacy. 2) It is efficient, since no single client in the network has to store data beyond their own. 3) With large enough networks, it is robust to some clients temporarily pausing updates to the model. However, while the standard collaborative learning framework performs well for the *average* client – returning a model trained on the clients' aggregated training data – optimizing its utility for the *individual* client is an ongoing challenge. Misalignment between the output of a collaborative learning system and the learning goals of individual clients can dis-incentivize participation, accounting in part for the slow adoption of collaborative learning in practice despite its powerful theoretical guarantees. The focus of this dissertation is designing collaborative learning systems that incentivize wide participation and guarantee good performance on diverse tasks. I examine three areas of potential misalignment between system-wide guarantees and individual client objectives and propose methods towards resolution. **I.** Clients will only want to join a collaborative learning system if the resulting model performs well on their individual tasks. We propose ways to personalize the global model to individual client learning objectives, and do so with optimal convergence guarantees. **II.** If the clients in the system are market competitors (e.g. autonomous car companies who are training self-driving models), a client may not want to participate in the system if doing so would advantage its competitors. Framing the collaborative learning system as an oligopoly of competitive firms, we show that its possible for competitors to collaborate and simultaneously increase their own revenues. **III.** Standard collaborative learning protocols guarantee a baseline level of privacy since they exchange models or gradients rather than raw data. However, extensive research has shown that these protections are insufficient, and in the real world, privacy concerns are a major deterrent for joining a collaborative learning system, especially if the clients have sensitive data (e.g. hospitals, banks). As a first step towards studying privacy dynamics in multi-agent systems, we design a game between potentially antagonistic agents (specifically buyers and a price-discriminating seller), and observe that an equilibrium arises in which the seller is incentivized to respect buyer privacy.

This dissertation consists of three chapters, each devoted to one of the following topics.

I: Personalized Collaborative Learning. The standard framework for collaborative (specifically federated) learning, FedAvg [1], averages clients' model weights, creating a global model which may perform badly for an individual client if there is heterogeneity across the clients' data distributions. Similarly, if individual clients perform impactful local updates in between global model updates, this can cause *client drift* [2], perturbing the final model away from the true optimum of the average loss function. An intuitive solution to this problem is to identify clusters of clients with similar data distributions and generate a model-per-cluster. Then clients who have similar learning objectives can benefit from collaboration without influence from dissimilar clients. Taking the ground-truth clustering to be clients who share a loss-function optimum ([3] show this is theoretically optimal), we iteratively learn the true clustering of clients while updating their models collaboratively via stochastic gradient descent. Our algorithm converges at an optimal rate for each individual client. In particular, the rate depends on three key parameters of the problem: the size of the client's cluster, the variance of the gradients within the client's cluster, and the dissimilarity of the client's loss optimum from the optima of other clusters. Additionally, we use a novel clustering method proposed in [4] which makes our algorithm provably robust against Byzantine attackers. We empirically verify our theory, showing that on learning tasks where personalization is critical our method out-performs existing personalization benchmarks and performs significantly better than standard collaborative learning algorithms.

II: Collaborative Learning among Competitors. Even when clients are competitors, there are still provable benefits to collaboration. To show this, we frame the collaborative learning system as an oligopoly in which multiple firms (the clients) compete to sell models to consumers (e.g. clients could be autonomous vehicle companies who are training models for their cars). It benefits consumers for the firms to collaborate and share their models with each other, especially if the firms' training distributions are complementary. However, firms have no incentive to collaborate if they lose revenue doing so. We propose a *defection-free* algorithm that allows firms to simultaneously collaborate and improve their revenue. In particular, the firms iteratively share their models with each other and update them in a way that guarantees no loss of revenue at any step. Even when the firms focus on improving their own revenues, we show that their model qualities converge to the Nash bargaining solution, thus optimizing for joint surplus.

III: Privacy Dynamics in Systems of Learning Agents. Privacy is a central area of concern in collaborative learning. While the standard collaborative setup ensures a light layer of privacy (clients send model weights or gradients to each other, not raw data), information about the clients' data distributions can still be inferred from the weights. This insight has inspired extensive work in cryptography and differential privacy seeking to more rigorously protect client information. We examine how incentives to protect privacy naturally arise from systems of learning agents. In particular, we look at a game between a seller and multiple buyers. Both the seller and buyers want to maximize their utility, and if buyers reveal their true types, the seller could price discriminate in an effort to increase utility. This in turn would incentivize buyers to lie about their true type. However, we show that over repeated interactions between the seller and buyers, an equilibrium is reached in which both sides of the market are incentivized to maintain a certain level of truthfulness and respect for privacy. In particular, we show that a utility-maximizing equilibrium is achieved in which the seller ignores buyer information (i.e. respects buyer privacy) with some probability. As a direction for future work, understanding how privacy protection endogenously arises from interactions between agents can help to formalize the privacy guarantees of collaborative learning systems.

Chapter 2

Personalized Collaborative Learning

This chapter is based on Werner et al. 2023, published in *Transactions on Machine Learning Research (2023)*

2.1 Introduction

We consider the federated learning setting in which there are N clients with individual loss functions $\{f_i\}_{i\in[N]}$ who seek to jointly train a model or multiple models. The defacto algorithm for problems in this setting is FedAvg [1] which has an objective of the form

$$x^*_{\text{FedAvg}} = \arg\min_{x\in\mathcal{X}} \frac{1}{N}\sum_{i\in[N]} f_i(x). \tag{2.1}$$

From (2.1), we see that FedAvg optimizes the average of the client losses. In many real-world cases however, clients' data distributions are heterogeneous, making such an approach unsuitable since the global optimum (2.1) may be very far from the optima of individual clients. Rather, we want algorithms which identify clusters of the clients that have relevant data for each other and that only perform training within each cluster. However, this is a challenging exercise since 1) it is unclear what it means for data distributions of two clients to be useful for each other, or 2) how to automatically identify such subsets without expensive multiple retraining [5]. In this work we propose algorithms which iteratively and simultaneously 1) identify K clusters amongst the clients by clustering their gradients and 2) optimize the clients' losses within each cluster.

Related Work

Personalization via Clustering. Personalization in federated learning has recently enjoyed tremendous attention (see [6, 7] for surveys). We focus on gradient-based clustering methods for personalized federated learning. Several recent works propose and analyze clustering methods. [8] alternately train a global model with FedAvg and partition clients into smaller clusters based on the global model's performance on their local data. [9] and [10] instead train personalized models from the start (as we do) without maintaining a global model. They iteratively update K models and, using empirical risk minimization, assign each of N clients one of the models at every step. In Section 2.2 we analyze these algorithms on constructed examples and in Section 2.3 compare them to our method.

Since our work is closest to [10], we highlight key similarities and differences. **Similarities**: 1) We both design stochastic gradient descent- and clustering-based algorithms for personalized federated learning. 2) We both assume sufficient intra-cluster closeness and inter-cluster separation of clients for the clustering task (Assumptions 1 and 2 in their work; Assumptions 4 and 5 in ours). 3) Our convergence rates both scale inversely with the number of clients and the inter-cluster separation parameter Δ. **Differences**: 1) They assume strong convexity of the clients' loss objectives, while our guarantees hold for all smooth (convex and non-convex) functions. 2) They cluster clients based on similarity of loss-function values whereas we cluster clients based on similarity of gradients. We show that clustering based on loss-function values instead of gradients can be overly sensitive to model initialization (see Fig. 2.1b). 3) Since we determine clusters based on distances in gradient space, we are able to apply an aggregation rule which makes our algorithm robust to some fraction of malicious clients. They determine cluster identity based on loss-function value and do not provide robustness guarantees.

Recently, [11] established lower bounds showing that the optimal strategy is to cluster clients who share the same optimum. Our algorithms and theoretical analysis are inspired by this lower-bound, and our gradient-based clustering approach makes our algorithms amenable to analysis à la their framework.

Multitask learning. Our work is closely related to multitask learning, which simultaneously trains separate models for different-but-related tasks. [12] and [13] both cast personalized federated learning as a multitask learning problem. In the first, the per-task models jointly minimize an objective that encodes relationships between the tasks. In the second, models are trained locally (for personalization) but regularized to be close to an optimal global model (for task-relatedness). These settings are quite similar to our setting. However, we use assumptions on gradient (dis)similarity across the domain space to encode relationships between tasks, and we do not maintain a global model.

Robustness. Our methods are provably robust in the Byzantine [14, 15] setting, where clients can make arbitrary updates to their gradients to corrupt the training process. Several works on Byzantine robust distributed optimization [15–19] propose aggregation rules in lieu of averaging as a step towards robustness. However, [20, 21] show that these rules are not in fact robust and perform poorly in practice. [22] are the first to provide a provably Byzantine-robust distributed optimization framework by combining a novel aggregation rule with momentum-based stochastic gradient descent. We use a version of their centeredclipping aggregation rule to update client gradients. Due to this overlap in aggregation rule, components of our convergence results are similar to their Theorem 6. However, our analysis is significantly complicated by our personalization and clustering structure. In particular, all non-malicious clients in [4, 22] have the same optimum and therefore can be viewed as comprising a single cluster, whereas we consider multiple clusters of clients (without necessarily assuming clients are i.i.d. within a cluster). The personalization algorithm in [13] also has robustness properties, but they are only demonstrated empirically and analyzed on toy examples.

Recent Empirical Approaches. Two recent works [23, 24] examine the setting in which clients' marginal distributions $p(x)$ differ, whereas most prior work only allows their conditional distributions $p(y|x)$ to differ. One of our experiments (Section 2.4) assumes heterogeneity between clients' marginal distributions, while the others (Section 2.4) assume heterogeneity only between their conditional distributions. In [23], the server maintains a global pool of modules (neural networks) from which clients, via a routing algorithm, efficiently select and combine sub-modules to create personalized models that perform well on their individual distributions. Extending the work of [25], [24] model each client's joint distribution as a mixture of Gaussian distributions, with the weights of the mixture personalized to each client. They then propose a federated Expectation-Maximization algorithm to optimize the parameters of the mixture model. In general, the contributions and style of our work and these others differ significantly. We focus on achieving and proving optimal theoretical convergence rates which we verify empirically, whereas [23] and [24] emphasize empirical application over theoretical analysis.

Our Contributions

To address the shortcomings in current approaches, we propose two personalized federated learning algorithms, which simultaneously cluster similar clients and optimize their loss objectives in a personalized manner. In each round of the procedure, we examine the client gradients to identify the cluster structure as well as to update the model parameters. Importantly, ours is the first method with theoretical guarantees for general non-convex loss functions, and not just restrictive toy settings. We show that our method enjoys both nearly optimal convergence, while also being robust to some malicious (Byzantine) client updates. This is again the first theoretical proof of the utility of personalization for Byzantine robustness. Specifically in this work,

- We show that existing or naive clustering methods for personalized learning, with stronger assumptions than ours, can fail in simple settings (Fig. 2.1).
- We design a robust clustering subroutine (Algorithm 3) whose performance improves with the separation between the cluster means and the number of data points being clustered. We prove nearly matching lower bounds showing its near-optimality (Theorem

2), and we show that the error due to malicious clients scales smoothly with the fraction of such clients (Theorem 1).

- We propose two personalized learning algorithms (Algorithm 2 and Algorithm 4) which converge at the optimal $\mathcal{O}(1/\sqrt{n_i T})$ rate in T for stochastic gradient descent for smooth non-convex functions and scale with n_i, the number of clients in client i's cluster.
- We empirically verify our theoretical assumptions and demonstrate experimentally that our learning algorithms benefit from collaboration, scale with the number of collaborators, are competitive with SOTA personalized federated learning algorithms, and are not sensitive to the model's initial weights (Section 2.4).

2.2 Existing Clustering Methods for Personalized Federated Learning

Our task at hand in this work is to simultaneously learn the clustering structure amongst clients and minimize their losses. Current methods do not rigorously check similarity of clients throughout the training process. Therefore they are not able to correct for early-on erroneous clustering (e.g. due to gradient stochasticity, model initialization, or the form of loss-functions far from their optima). In the next section we demonstrate such failure modes of existing algorithms.

Failure Modes of Existing Methods

The first algorithm we discuss, Myopic-Clustering, does not appear in the existing literature, but we create it in order to motivate the design of our method (Algorithm 2). In particular, it is a natural first step towards our method, but has limitations which we correct when designing our algorithm.

Myopic-Clustering (Algorithm 1). At every step, each client computes their gradient at their current model and sends the gradient to a central server. The server clusters the gradients and sends each cluster center to the clients assigned to that cluster. Each client then performs a gradient descent update on their model with their received cluster center. This is a natural federated clustering procedure and it is communication-efficient ($\mathcal{O}(N)$). However, it has two issues: 1) If it makes a clustering mistake at one step, models will be updated with the wrong set of clients. This can cause models to diverge from their optima, gradients of clients in the same cluster to drift apart, and gradients of clients in different clusters to drift together, thus obscuring the correct clustering going forward. Furthermore, these errors can compound over rounds. 2) Even if Myopic-Clustering clusters clients perfectly at each step, the clients' gradients will approach zero as the models converge to their optima. This

Algorithm 1 Myopic-Clustering

```
Input Learning rate: $\eta$. Initial parameters: $\{x_{1,0} = ... = x_{N,0} = x_0\}$.
for round $t \in [T]$ do
    for client $i$ in [N] do
        Client $i$ sends $g_i(x_{i,t-1})$ to server.
        Server clusters $\{g_i(x_{i,t-1})\}_{i\in[N]}$, generating cluster centers $\{v_{k,t}\}_{k\in[K]}$.
        Server sends $v_{k_i,t}$ to client $i$, where $k_i$ denotes the cluster to which client $i$ is assigned.
        Client $i$ computes update: $x_{i,t} = x_{i,t-1} - \eta v_{k_i,t}$.
Output: Personalized parameters: $\{x_{1,T}, ..., x_{N,T}\}$.
```

means that clients from different clusters will appear to belong to the same cluster as the algorithm converges and all clients will collapse into a single cluster.

The following example (Fig. 2.1a) demonstrates these failure modes of Myopic-Clustering.

Let $N = 3$ and $K = 2$, with client loss functions

$$
\begin{aligned}
f_1(x) &= \frac{1}{6\eta}x^2 \\
f_2(x) &= \begin{cases} 4(x-1)^3 + 3(x-1)^4 + 1 & x < 1 \\ \frac{1}{2\eta}(x-1)^2 + 1 & x \geq 1, \end{cases} \\
f_3(x) &= \frac{1}{2\eta}(x-2)^2,
\end{aligned}
$$

where η is the learning rate of the algorithm. With this structure, clients $\{1, 2\}$ share the same global minimum and belong to the same cluster, and client $\{3\}$ belongs to its own cluster. Suppose Myopic-Clustering is initialized at $x_0 = 1.5$. At step 1, the client gradients computed at $x_0 = 1.5$ are $1/2\eta$, $1/2\eta$, and $-1/2\eta$ respectively. Therefore, clients $\{1, 2\}$ are correctly clustered together and client $\{3\}$ alone at this step. After updates, the clients' parameters will next be $x_{1,1} = 1, x_{2,1} = 1, x_{3,1} = 2$ respectively. At this point, clients $\{2, 3\}$ will be incorrectly clustered together since their gradients will both be 0, while client $\{1\}$ will be clustered alone. As the algorithm proceeds, clients $\{2, 3\}$ will always be clustered together and will remain at $x = 1$ and $x = 2$ respectively, while client $\{1\}$ will converge to its optimum at $x = 0$. Consequently, two undesirable things happen: 1) Client $\{2\}$ gets stuck at the saddle point at $x = 1$ which occurred when it was incorrectly clustered with client $\{3\}$ at $t = 1$ and subsequently did not recover. 2) All gradients converge to 0, so at the end of the algorithm all clients are clustered together.

To further motivate the design choice for our algorithms, we now discuss three clustering-based algorithms in the literature on personalized federated learning. In particular, we generate counter-examples on which they fail and show how our algorithm avoids such pitfalls.

The first two algorithms IFCA [10] and HypCluster [9] are closely related. They both cluster loss function values rather than gradients, and like our algorithm they avoid the myopic nature of Myopic-Clustering by, at each step, computing all client losses at all current cluster parameters to determine the clustering. However, as we show in the next example (Fig. 2.1b), they are brittle and sensitive to initialization.

IFCA [10]. Let $N = 2$, $K = 2$ with loss functions

$$f_1(x) = (x + 0.5)^2$$
$$f_2(x) = (x - 0.5)^2,$$

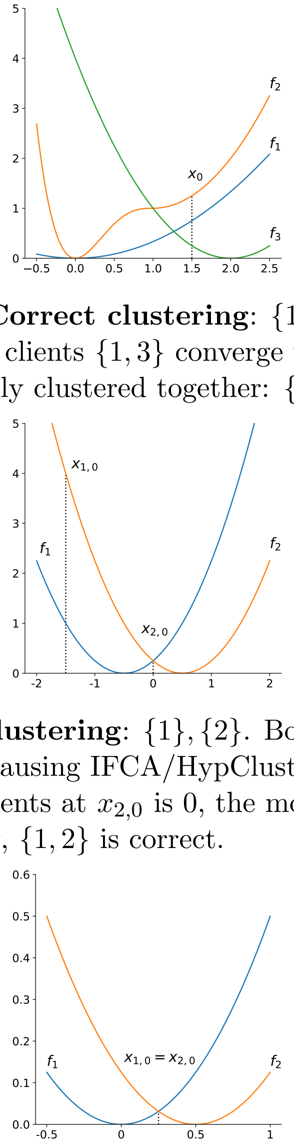
and initialize clusters 1 and 2 at $x_{1,0} = -1.5$ and $x_{2,0} = 0$ respectively. Given this setup, both clients initially select cluster 2 since their losses at $x_{2,0}$ are smaller than at $x_{1,0}$.

Option I: At $x_{2,0} = 0$, the client gradients will average to 0. Consequently the models will remain stuck at their initializations, and both clients will be incorrectly assigned to cluster 2.

Option II: Both clients individually run τ steps of gradient descent starting at their selected model $x_{2,0}$ (i.e. perform the `LocalUpdate` function in line 18 of IFCA). Since the clients' individually updated models will be symmetric around 0 after this process, the server will compute cluster 2's model update in line 15 of IFCA as: $x_{2,1} \leftarrow 0 = x_{2,0}$. Consequently, the outcome is the same as in Option I: the models never update and both clients are incorrectly assigned to cluster 2.

HypCluster [9]. This algorithm is a centralized version of Option II of IFCA. The server alternately clusters clients by loss function value and runs stochastic gradient descent per-cluster using the clients' data. It performs as Option II of IFCA on the example above.

Finally, we discuss Clustered Federated Learning, the algorithm proposed in [8], which runs the risk of clustering too finely, as in the next example (Fig. 2.1c).



(a) **Myopic-Clustering** ($\eta = 0.5$). **Correct clustering**: $\{1, 2\}$ and $\{3\}$. Client $\{2\}$ gets stuck at $x = 1$, not reaching its optimum, and clients $\{1, 3\}$ converge to their optima. All gradients being 0 at this point, the clients are incorrectly clustered together: $\{1, 2, 3\}$.

(b) **IFCA/HypCluster**. **Correct clustering**: $\{1\}, \{2\}$. Both clients' function values are smaller at initialization point $x_{2,0}$ than $x_{1,0}$ causing IFCA/HypCluster to initially cluster them together. Since the average of the clients' gradients at $x_{2,0}$ is 0, the models never update and the algorithm thinks the initial erroneous clustering, $\{1, 2\}$ is correct.

(c) **Clustered Federated Learning**. **Correct clustering:** $\{1\}$**,** $\{2, 3\}$ (client 3 not drawn due to its stochastic gradient – details on pg. 5). Clustered FL averages gradients of clients $\{1\}$ and $\{2\}$ to 0, clustering them together, and with non-0 probability clusters $\{3\}$ separately due to its stochastic gradient. Based on this initial erroneous, the algorithm partitions the clients $\{1, 2\}$ and $\{3\}$ and recursively runs on each group, never recovering the correct clustering.

Figure 2.1: We show how existing personalized FL algorithms miscluster and fail to converge on constructed examples.

Clustered Federated Learning [8]**.** Clustered Federated Learning operates by recursively bi-partitioning the set of clients based on the clients' gradient values at the FedAvg optimum.

Consider the following example. Let $N = 3$ and $K = 2$ with client gradients

$$
\begin{aligned}
g_1(x) &= x \\
g_2(x) &= x - {}^1\!/_2 \\
g_3(x) &= \begin{cases} x & \text{w.p. } {}^1\!/_2 \\ x - 1 & \text{w.p. } {}^1\!/_2. \end{cases}
\end{aligned}
$$

Therefore the correct clustering here is $\{1\}$ and $\{2, 3\}$. The FedAvg optimum is $x^*_{\text{FedAvg}} = {}^1\!/_4$, at which the clients' gradient values are $g_1({}^1\!/_4) = {}^1\!/_4$, $g_2({}^1\!/_4) = -{}^1\!/_4$ and $g_3 = {}^1\!/_4$ w.p. ${}^1\!/_2$. Based on this computation, Clustered Federated Learning partitions the client set into $\{1, 2\}$ and $\{3\}$ w.p. ${}^1\!/_2$ and then proceeds to run the algorithm separately on each sub-cluster. Therefore, the algorithm never corrects its initial error in separating clients $\{2\}$ and $\{3\}$.

The behaviour of these algorithms motivates our method Federated-Clustering, which by rigorously checking client similarity at every step of the training process can recover from past clustering errors.

2.3 Proposed Method: Federated-Clustering

At a high level, Federated-Clustering works as follows. Each client i maintains a personalized model which, at every step, it broadcasts to the other clients $j \neq i$. Then each client j computes its gradient on clients i's model parameters and sends the gradient to client i. Finally, client i runs a clustering procedure on the received gradients, determines which other clients have gradients closest to its own at its current model, and updates its current model by averaging the gradients of these similar clients. By the end of the algorithm, ideally each client has a model which has been trained only on the data of similar clients.

The core of Federated-Clustering is a clustering procedure, Threshold-Clustering (Algorithm 3), which identifies clients with similar gradients at each step. This clustering procedure, which we discuss in the next section, has two important properties: it is robust and its error rate is near-optimal.

Notation. For an arbitrary integer N, we let $[N] = \{1, ..., N\}$. We take $a \gtrsim b$ to mean there is a sufficiently large constant c such that $a \geq cb$, $a \lesssim b$ to mean there is a sufficiently small constant c such that $a \leq cb$, and $a \approx b$ to mean there is a constant c such that $a = cb$. We write $i \sim j$ if clients i and j belong to the same cluster, $i \overset{i.i.d.}{\sim} j$ if they belong to the same cluster and their data is drawn independently from identical distributions (we will sometimes equivalently write $z_i \overset{i.i.d.}{\sim} z_j$, where z_i and z_j are arbitrary points drawn from clients i's and j's distributions), and $i \not\sim j$ if they belong to different clusters. For two different clients i and j, same cluster or not, we write $i \neq j$. Finally, n_i denotes the number of clients in client i's cluster, $\delta_i = {}^{n_i}\!/_N$ denotes the fraction of clients in client i's cluster, and β_i denotes the fraction of clients that are malicious from client i's perspective.

Analysis of Clustering Procedure

Given the task of clustering N points into K clusters, at step l our clustering procedure has current estimates of the K cluster-centers, $v_{1,l}, ..., v_{K,l}$. To update each estimate $v_{k,l+1} \leftarrow v_{k,l}$, it constructs a ball of radius $\tau_{k,l}$ around $v_{k,l}$. If a point falls inside the ball, the point retains its value; if it falls outside the ball, its value is mapped to the current cluster-center estimate. The values of all the points are then averaged to set $v_{k,l+1}$ (update rule (2.4)). The advantage of this rule is that it is very conservative. If our algorithm is confident that its current cluster-center estimate is close to the true cluster mean (i.e. there are many points nearby), it will confidently improve its estimate by taking a large step in the right direction (where the step size and direction are determined mainly by the nearby points). If our algorithm is not confident about being close to the cluster mean, it will tentatively improve its estimate by taking a small step in the right direction (where the step size and direction are small since the majority of points are far away and thus do not change the current estimate).

To analyze the theoretical properties of this procedure, we look at a natural setting in which clients within the same cluster have i.i.d. data (for analysis of our federated learning algorithm, we will relax this strong notion of intra-cluster similarity). In particular, in our setting there are N points $\{z_1, ..., z_N\}$ which can be partitioned into K clusters within which points are i.i.d.. We assume the following.

- **Assumption 1** (Intra-cluster Similarity): For all $i \sim j$,

$$z_i \overset{i.i.d.}{\sim} z_j.$$

- **Assumption 2** (Inter-cluster Separation): For all $i \nsim j$,

$$\|\mathbb{E}z_i - \mathbb{E}z_j\|^2 \geq \Delta^2.$$

- **Assumption 3** (Bounded Variance): For all z_i,

$$\mathbb{E}\|z_i - \mathbb{E}z_i\|^2 \leq \sigma^2.$$

Theorem 1. *Suppose there N points $\{z_i\}_{i\in[N]}$ for which Assumptions [1-3] hold with inter-cluster separation parameter $\Delta \gtrsim \sigma/\delta_i$. Running Algorithm 3 for*

$$l \gtrsim \max\left\{1, \max_{i\in[N]} \frac{\log(\sigma/\Delta)}{\log(1 - \delta_i/2)}\right\}$$

steps with fraction of malicious clients $\beta_i \lesssim \delta_i$ and thresholding radius $\tau \approx \sqrt{\delta_i \sigma \Delta}$ guarantees that

$$\mathbb{E}\|v_{k_i,l} - \mathbb{E}z_i\|^2 \lesssim \frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \beta_i \sigma \Delta. \tag{2.2}$$

Proof. See 2.6. □

Supposing $\beta_i = 0$, if we knew the identity of all points within z_i's cluster, we would simply take their mean as the cluster-center estimate, incurring estimation error of ${}^{\sigma^2}/_{n_i}$ (i.e. the sample-mean's variance). Since we don't know the identity of points within clusters, the additional factor of ${}^{\sigma^3}/_{\Delta}$ in (2.2) is the price we pay to learn the clusters. This additional term scales with the difficulty of the clustering problem. If true clusters are well-separated and/or the variance of the points within each cluster is small (i.e. Δ is large, σ^2 is small), then the clustering problem is easier and our bound is tighter. If clusters are less-well-separated and/or the variance of the points within each cluster is large, accurate clustering is more difficult and our bound weakens.

Setting τ. To achieve the rate in (2.2), we set $\tau \approx \sqrt{\sigma\Delta}$, which is the geometric mean of the standard deviation, σ, of points belonging to the same cluster and the distance, Δ, to a different cluster. The intuition for this choice is that we want the radius for each cluster to be at least as large as the standard deviation of the points belonging to that cluster in order to capture in-cluster points. The radius could be significantly larger than the standard deviation if Δ is large, thus capturing many non-cluster points as well. However, the conservative nature of our update rule (2.4) offsets this risk. By only updating the center with a step-size proportional to the *fraction* of points inside the ball, it limits the influence of any mistakenly captured points.

Threshold-Clustering has two important properties which we now discuss: it is Byzantine robust and has a near-optimal error rate.

Robustness

We construct the following definition to characterize the robustness of Algorithm 3.

Definition 1 (Robustness). *An algorithm $\mathcal{A}$ is* robust *if the error introduced by bad clients can be bounded i.e. malicious clients do not have an arbitrarily large effect on the convergence. Specifically, for a specific objective, let $\mathcal{E}_1$ be the base error of $\mathcal{A}$ with no bad clients, let β be the fraction of bad clients, and let $\mathcal{E}_2$ be some bounded error added by the bad points. Then $\mathcal{A}$ is* robust *if*

$$Err(\mathcal{A}) \leq \mathcal{E}_1 + \beta\mathcal{E}_2.$$

Threat model. Our clustering procedure first estimates the centers of the K clusters from the N points and constructs a ball of radius τ_k around the estimated center of each cluster k. If a point falls inside the ball, the point retains its value; if it falls outside the ball, its value is mapped to the current cluster-center estimate. Following the update rule (2.4), the values of all the points are averaged to update the cluster-center estimate. Therefore, a bad point that wants to distort the estimate of the k'th cluster's center has the most influence by placing itself just within the boundary of the ball around that cluster-center, i.e. at τ_k-distance from the cluster-center.

From (2.2) we see that the base squared-error of Algorithm 3 in estimating z_i's cluster-center is $\lesssim {}^{\sigma^2}/_{n_i} + {}^{\sigma^3}/_{\Delta}$, and that the bad points introduce extra squared-error of order $\sigma\Delta$. Given our threat model, this is exactly expected. The radius around z_i's cluster-center is order $\sqrt{\sigma\Delta}$. Therefore, bad points placing themselves at the edge of the ball around z_i's cluster-center estimate will be able to distort the estimate by order $\sigma\Delta$. The scaling of this extra error by β_i satisfies our definition of robustness, and the error smoothly vanishes as $\beta_i \to 0$.

Near-Optimality

The next result shows that the upper bound (2.2) on the estimation error of Algorithm 3 nearly matches the best-achievable lower bound. In particular, it is tight within a factor of ${}^{\sigma}/_{\Delta}$.

Theorem 2 (Near-optimality of **Threshold-Clustering**)**.** *For any algorithm $\mathcal{A}$, there exists a mixture of distributions $\mathcal{D}_1 = (\mu_1, \sigma^2)$ and $\mathcal{D}_2 = (\mu_2, \sigma^2)$ with $\|\mu_1 - \mu_2\| \geq \Delta$ such that the estimator $\hat{\mu}_1$ produced by $\mathcal{A}$ has error*

$$\mathbb{E}\|\hat{\mu}_1 - \mu_1\|^2 \geq \Omega\left(\frac{\sigma^4}{\Delta^2} + \frac{\sigma^2}{n_i}\right).$$

Proof. See 2.6. □

Federated-Clustering on Examples in Section 2.2

We describe how Federated-Clustering successfully handles the examples in Section 2.2.

Example 1: Fig. 2.1a. Federated-Clustering checks at every step the gradient values of all N clients at the current parameters of all K clusters. This verification process avoids the type of errors made by Myopic-Clustering. For instance, at $t = 1$ when Myopic-Clustering makes its error, Federated-Clustering computes the gradients of all clients at client $\{1\}$'s current parameters: $g_1(1) = {}^{1}/_{3\eta}$, $g_2(1) = 0$, and $g_3(1) = -{}^{1}/_{\eta}$. Therefore it correctly clusters $\{1, 2\}$ together at this point, and client $\{2\}$'s parameters update beyond the saddle-point and converge to the global minimum at $x = 0$.

Example 2: Fig. 2.1b. By clustering clients based on gradient instead of loss value, Federated-Clustering initially computes the clients' gradients of $+1$ and -1 respectively at $x_{2,0} = 0$, and given the continued separation of their gradients around 0 as the algorithm converges, correctly identifies that they belong to different clusters.

Example 3: Fig. 2.1c. Recall how Clustered FL fails on this example. Based on an initial clustering error, it partitions the clients incorrectly early on and then evaluates each subset separately going forward, thus never recovering the correct clustering. Our algorithm avoids this type of mistake by considering all clients during each clustering at every step.

Analysis of Federated-Clustering

We now proceed with the analysis of Federated-Clustering. First, we establish necessary assumptions: intra-cluster similarity, inter-cluster separation, bounded variance of stochastic gradients, and smoothness of loss objectives.

- **Assumption 4** (Intra-cluster Similarity): For all x, $i \sim j$, and some constant $A \geq 0$,

  $$\|\nabla f_i(x) - \nabla \bar{f}_i(x)\|^2 \leq A^2 \|\nabla \bar{f}_i(x)\|^2,$$

  where $\bar{f}_i(x) \triangleq \frac{1}{n_i} \sum_{j \sim i} f_j(x)$.

- **Assumption 5** (Inter-cluster Separation): For all x, $i \not\sim j$, and some constant $D \geq 0$,

  $$\|\nabla f_i(x) - \nabla f_j(x)\|^2 \geq \Delta^2 - D^2 \|\nabla f_i(x)\|^2.$$

This formulation is motivated by the information theoretic lower-bounds of [11] who show that the optimal clustering strategy is to group all clients with the same optimum (even if they are non-iid). Assumptions 4 and 5 are in fact a slight strengthening of this very statement. To see this, note that for a client with loss function $f_i(x)$, belonging to cluster $\bar{f}_i(x)$ with a first-order stationary points $\bar{x}^*$, Assumption 4 implies that if $\nabla \bar{f}_i(\bar{x}^*) = 0 \Rightarrow \nabla f_i(\bar{x}^*) = 0$, and so $\bar{x}^*$ is also a stationary point for client i. Thus, all clients within a cluster have shared stationary points. Assumption 4 further implies that the gradient difference elsewhere away from the optima is also bounded. This latter strengthening is motivated by the fact the the loss functions are smooth, and so the gradients cannot diverge arbitrarily as we move away from the shared optima. In fact, it is closely related to the strong growth condition (equation (1) in [26]), which is shown to be a very useful notion in practical deep learning. We also empirically verify its validity in Fig. 2.2 (left).

Similarly, Assumption 5 is a strengthening of the condition that clients across different clusters need to have different optima. For two clients i and j who belong to different clusters and with first-order stationary points x_i^* and x_j^*, Assumption 5 implies that $\|\nabla f_i(x_j^*)\|^2 \geq \Delta$ and $\|\nabla f_j(x_i^*)\|^2 \geq \Delta$. Thus, they do not share any common optimum. Similar to the Assumption 4, Assumption 5 also describes what happens elsewhere away from the optima - it allows for the difference between the gradients to be smaller than Δ as we move away from the stationary points. Again, this specific formulation is motivated by smoothness of the loss function, and empirical validation (Fig. 2.2, right).

In the following lemma, we give a specific setting in which Assumptions 4 and 5 hold.

Lemma 1. *Suppose losses f_i are L-smooth and μ-strongly convex and clients in the same cluster have the same optima. Then for all clients i*

$$\|\nabla f_i(x) - \nabla \bar{f}_i(x)\|^2 \leq ({}^{2L}/_{\mu})^2 \|\nabla \bar{f}_i(x)\|^2,$$

and for clients $i \not\sim j$ in different clusters

$$\|\nabla f_i(x) - \nabla f_j(x)\|^2 \geq \frac{1}{2}(\max_{j \not\sim i} \|\nabla f_j(x_i^*)\|)^2 - 2(1 + ({}^{L}/_{\mu})^2)\|\nabla f_i(x)\|^2$$

for all x.

Proof. See 2.6. □

- **Assumption 6** (Bounded Variance of Stochastic Gradients): For all x,

  $$\mathbb{E}\|g_i(x) - \nabla f_i(x)\|^2 \leq \sigma^2,$$

  where $\mathbb{E}[g_i(x)|x] = \nabla f_i(x)$ and each client i's stochastic gradients $g_i(x)$ are independent.

- **Assumption 7** (Smoothness of Loss Functions): For any x, y,

  $$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|.$$

Theorem 3. *Let* Assumptions [4-7] *hold with inter-cluster separation parameter*

$$\Delta \gtrsim {}^{\max(1,A^4)\max(1,D^2)\sigma}/_{\delta_i}.$$

Under these conditions, suppose we run Algorithm 2 for T rounds with learning rate $\eta \leq {}^1/_L$, fraction of malicious clients $\beta_i \lesssim \delta_i$, and batch size

$$|B_i| \gtrsim \min(\sqrt{\max(1, A^2)({}^{\sigma^2}/_{n_i} + {}^{\sigma^3}/_{\Delta} + \beta_i \sigma \Delta) m_i}, m_i),$$

where m_i is the size of client i's training dataset. If, in each round $t \in [T]$, we cluster with radius $\tau \approx \sqrt{\delta_i \sigma \Delta}$ for

$$l \geq \max\left\{1, \max_{i \in N]} \frac{\log({}^{\sigma}/_{\sqrt{|B_i|\Delta}})}{\log(1 - {}^{\delta_i}/_{2})}\right\}$$

steps, then

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2 \lesssim \sqrt{\frac{\max(1, A^2)({}^{\sigma^2}/_{n_i} + {}^{\sigma^3}/_{\Delta} + \beta_i \sigma \Delta)}{T}}. \tag{2.3}$$

Proof. See 2.6. □

We note a few things. 1) The rate in (2.3) is the optimal rate in T for stochastic gradient descent on non-convex functions [27]. 2) The dependence on $\sqrt{\sigma^2/n_i}$ is intuitive, since convergence error should increase as the variance of points in the cluster increases and decrease as the number of points in the cluster increases. It is also optimal as shown in [11]. 3) The dependence on $\sqrt{\beta_i \sigma \Delta}$ is also expected. We choose a radius $\tau \approx \sqrt{\sigma \Delta}$ for clustering. Given our threat model, the most adversarial behavior of the bad clients from client i's perspective is to place themselves at the edge of the ball surrounding the estimated location of i's gradient, thus adding error of order $\sqrt{\beta_i \sigma \Delta}$. When there are no malicious clients, this extra error vanishes. 4) If the constraint on batch-size in Theorem 3 requires $|B_i| = m_i$, then the variance of stochastic gradients vanishes and the standard $\mathcal{O}(1/T)$ rate for deterministic gradient descent is recovered (see equation (2.22) in proof). We also note that as long as there are no malicious clients (i.e. $\beta_i \sigma \Delta$ term is 0), there are a large enough number of clients n_i in the cluster, and inter-cluster separation Δ is sufficiently larger than the inter-cluster-variance σ^2, then minimum-batch size will likely be less than m_i.

If losses are smooth and strongly convex, our proof techniques for Theorem 3 get an $\mathcal{O}(1/T)$ convergence rate with the specific constants A, D, and Δ stated in Lemma 1.

Privacy. Since Federated-Clustering requires clients to compute distances between gradients, they must share their models and gradients which compromises privacy. The focus of our work is not on optimizing privacy, so we accommodate only the lightest layer of privacy for federated learning: sharing of models and gradients rather than raw data. Applying more robust privacy techniques is a direction for future work. In the meantime, we refer the reader to the extensive literature on differential privacy, multi-party computation, and homomorphic encryption in federated learning.

Communication Overhead. At each step, Federated-Clustering requires $\mathcal{O}(N^2)$ rounds of communication since each client sends its model to every other client, evaluates its own gradient at every other client's model and then sends this gradient back to the client who owns the model. We pay this communication price to mitigate the effect of past clustering mistakes. For example, say at one round a client mis-clusters itself and updates its model incorrectly. At the next step, due to communication with all other clients, it can check the gradients of all other clients at its current model, have a chance to cluster correctly at this step, and update its model towards the optimum, regardless of the previous clustering error. Recall on the other hand that an algorithm like Myopic-Clustering (Alg. 1), while communication efficient (N rounds per step), may not recover from past clustering mistakes since it doesn't check gradients rigorously in the same way. In the next section, we propose a more communication-efficient algorithm, Momentum-Clustering (Algorithm 4), which clusters momentums instead of gradients (reducing variance and thus clustering error) and requires only $\mathcal{O}(N)$ communication rounds per step.

Algorithm 2 Federated-Clustering

```
Input Learning rate: $\eta$. Initial parameters for each client: $\{x_{1,0}, ..., x_{N,0}\}$. Batch-size $|B_i|$ (see Theorem 3 for a lower bound on this quantity)[1]
for client $i \in [N]$ do
    Send $x_{i,0}$ to all clients $j \neq i$.
for round $t \in [T]$ do
    for client $i$ in [N] do
        Compute $g_i(x_{j,t-1})$ with batch-size $|B_i|$[2]and send to client $j$ for all $j \neq i \in [N]$.
        Compute $v_{i,t} \leftarrow \texttt{Threshold-Clustering}(\{g_j(x_{i,t-1})\}_{j\in[N]}; 1 \text{ cluster}; g_i(x_{i,t-1}))$.
        Update parameter: $x_{i,t} = x_{i,t-1} - \eta v_{i,t}$.
        Send $x_{i,t}$ to all clients $j \neq i$.
Output: Personalized parameters: $\{x_{1,T}, ..., x_{N,T}\}$.
```

Algorithm 3 Threshold-Clustering

```
Input Points to be clustered: $\{z_1, ..., z_N\}$. Number of clusters: $K$. Cluster-center initializations: $\{v_{1,0}, ..., v_{K,0}\}$.
for round $l \in [M]$ do
    for cluster $k$ in [K] do
        Set radius $\tau_{k,l}$.
        Update cluster-center estimate:
            $$v_{k,l} = \frac{1}{N}\sum_{i=1}^{N}\Big(z_i\mathbb{1}(\|z_i - v_{k,l-1}\| \leq \tau_{k,l}) + v_{k,l-1}\mathbb{1}(\|z_i - v_{k,l-1}\| > \tau_{k,l})\Big). \qquad (2.4)$$
Output: Cluster-center estimates $\{v_1 = v_{1,M}, ..., v_K = v_{K,M}\}$.
```

Improving Communication Overhead with Momentum

Federated-Clustering is inefficient, requiring N^2 rounds of communication between clients at each step (each client computes their gradient at every other client's parameter). Since momentums change much more slowly from round-to-round than gradients, a past clustering mistake will not have as much of a harmful impact on future correct clustering and convergence as when clustering gradients.

In Algorithm 4, at each step each client computes their momentum and sends it to the server. The server clusters the N momentums, computes an update per-cluster, and sends

[1] The batch-size constraint reduces variance of the stochastic gradients (Lemma 7). In Section 2.3 we propose another algorithm Momentum-Clustering for which there is no batch-size restriction and which reduces variance by clustering momentums instead of gradients.

[2] $g_i(x_{j,t-1}) = \frac{1}{|B_i|}\sum_b g_i(x_{j,t-1}; b)$, where $g_i(x_{j,t-1}; b)$ is the gradient computed using sample b in the batch.

the update to the clients in each cluster. Therefore, communication is limited to N rounds per step.

Analysis of Momentum-Clustering

The analysis of the momentum based method requires adapting the intra-cluster similarity and inter-cluster separation assumptions from before.

- **Assumption 8** (Intra-cluster Similarity): For all $i \sim j$ and $t \in [T]$,

$$m_{i,t} \overset{i.i.d.}{\sim} m_{j,t},$$

  where $m_{i,t}$ is defined as in (2.6).

- **Assumption 9** (Inter-cluster Separation): For all $i \not\sim j$ and $t \in [T]$,

$$\|\mathbb{E}m_{i,t} - \mathbb{E}m_{j,t}\|^2 \geq \Delta^2.$$

Note that the intra-cluster similarity assumption in this momentum setting is stronger than in the gradient setting (Assumption 4): namely we require that the momentum of clients in the same cluster be i.i.d. at all points. This stronger assumption is the price we pay for a simpler and more practical algorithm. Finally, due to the fact that momentums are low-variance counterparts of gradients (Lemma 14), we can eliminate constraints on the batch size and still achieve the same rate.

Theorem 4. *Let* Assumptions [6-9] *hold with inter-cluster separation parameter* $\Delta \gtrsim {}^{\sigma}/_{\delta_i}$. *Under these conditions, suppose we run Algorithm 4 for* T *rounds with learning rate* $\eta \lesssim \min\left\{\frac{1}{L}, \sqrt{\frac{\mathbb{E}(f_i(x_{i,0}) - f_i^*)}{LT({}^{\sigma^2}/_{n_i} + {}^{\sigma^3}/_{\Delta})}}\right\}$ *(*f_i^* *is the global minimum of* f_i*), fraction of bad clients* $\beta_i \lesssim \delta_i$*, and momentum parameter* $\alpha \gtrsim L\eta$*. If, in each round* $t \in [T]$*, we cluster with radius* $\tau \approx \sqrt{\delta_i \sigma \Delta}$ *for*

$$l \geq \max\left\{1, \max_{i \in N]} \frac{\log({}^{\sqrt{\alpha}\sigma}/_{\Delta})}{\log(1 - {}^{\delta_i}/_{2})}\right\}$$

steps, then for all $i \in [N]$

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2 \lesssim \sqrt{\frac{{}^{\sigma^2}/_{n_i} + {}^{\sigma^3}/_{\Delta}}{T}} + \frac{\beta_i \sigma \Delta}{T^{\frac{1}{4}}({}^{\sigma^2}/_{n_i} + {}^{\sigma^3}/_{\Delta})^{\frac{1}{4}}}. \tag{2.5}$$

Proof. See 2.6. □

We see from (2.5) that when there are no malicious clients ($\beta_i = 0$), Momentum-Clustering achieves the same $\sqrt{{}^{\sigma^2}/_{n_i T}}$ convergence rate observed in (2.3), with no restrictions on the batch size.

Algorithm 4 Momentum-Clustering

```
Input Learning rate: $\eta$. Initial parameters: $\{x_{1,0} = ... = x_{N,0}\}$.
for round $t \in [T]$ do
    for client $i$ in [N] do
        Client $i$ sends
            $$m_{i,t} = \alpha g_i(x_{i,t-1}) + (1-\alpha)m_{i,t-1} \qquad (2.6)$$
        to server.
        Server generates cluster centers
            $$\{v_{k,t}\}_{k\in[K]} \leftarrow \texttt{Threshold-Clustering}(\{m_{i,t}\}; K \text{ clusters}; \{v_{k,t-1}\}_{k\in[K]})$$
        and sends $v_{k_i,t}$ to client $i$, where $k_i$ denotes the cluster to which $i$ is assigned in this step.
        Client $i$ computes update: $x_{i,t} = x_{i,t-1} - \eta v_{k_i,t}$.
Output: Personalized parameters: $\{x_{1,T}, ..., x_{N,T}\}$.
```

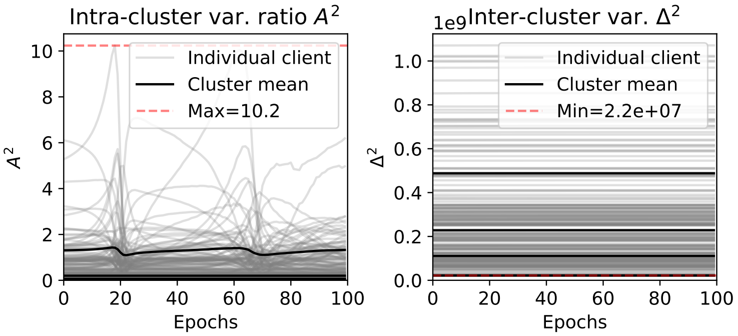


Figure 2.2: Here, we show empirically on a synthetic dataset that the intra-cluster variance ratio is upper-bounded by a constant (left subplot) and the inter-cluster variance that is lower-bounded by a constant (right subplot).

2.4 Experiments

In this section, we first use a synthetic dataset to verify the assumptions and rates claimed in our theoretical analysis in the previous section; and second, we use the MNIST dataset [28] and CIFAR dataset [29] to compare our proposed algorithm, Federated-Clustering, with existing state-of-the-art federated learning algorithms. All algorithms are implemented with PyTorch [30], and code for all experiments is available at this github repo.

Synthetic dataset

Construction of synthetic dataset. We consider a synthetic linear regression task with squared loss for which we construct $K = 4$ clusters, each with $n_i = 75$ clients. Clients in cluster $k \in [K]$ share the same minimizer $x_k^\star \in \mathbb{R}^d$. For each client i in cluster k, we generate a sample matrix $A_i \in \mathbb{R}^{d\times n}$ from $\mathcal{N}(k, \mathbf{1}_{d\times n})$ and compute the associated target as $y_i = A_i^\top x_k^\star \in \mathbb{R}^n$. We choose the model dimension $d = 10$ to be greater than the number of local samples $n = 9$ such that the local linear system $y_i = A_i^\top x$ is overdetermined and the error $\|x^\star - x\|_2^2$ is large. A desired federated clustering algorithm determines the minimizer by incorporating information from other clients $j \sim i$ in the same cluster.

Estimating constants in Assumptions 4 and 5. In Fig. 2.2, using the above synthetic dataset

- we estimate the intra-cluster variance ratio A^2 by finding the upper bound of (2.7)

$$\frac{\|\nabla f_i(x) - \nabla \bar{f}_i(x)\|_2^2}{\|\nabla \bar{f}_i(x)\|_2^2}; \tag{2.7}$$

- we estimate the inter-cluster variance Δ^2 by setting $D = 0$ and computing the lower bound of (2.8)

$$\|\nabla f_i(x) - \nabla f_j(x)\|_2^2. \tag{2.8}$$

We run Federated-Clustering with perfect clustering assignments and estimate these bounds over time. The result is shown in Fig. 2.2 where grey lines are the quantities in (2.7) and (2.8) for individual clients, black lines are those quantities averaged within clusters, and red dashed lines are empirical bounds on the quantities. The left figure demonstrates that the intra-cluster variance ratio does not grow with time and can therefore reasonably be upper bounded by a constant A^2. Similarly, the right figure shows that the inter-cluster variance can be reasonably lower bounded by a positive constant Δ^2. These figures/prfl empirically demonstrate that Assumptions 4 and 5 are realistic in practice.

Performance. In Fig. 2.3, we compare the performance of our algorithm Federated-Clustering (FC) with several baselines: standalone training (Local), IFCA [10], FedAvg (Global) [1], and distributed training with ground truth (GT) cluster information. We consider the synthetic dataset from before, starting with cluster parameters $(K, n_i) = (4, 4)$ and observe performance when increasing parameters to $n_i = 16$ and $K = 16$ separately. In each step of optimization, we run Threshold-Clustering for $l = 10$ rounds so that heuristically the outputs are close enough to cluster centers, cf. Fig. 2.7. We tune the learning rate separately for each algorithm through grid search, but preserve all other algorithmic setups. Our algorithm outperforms the non-oracle baselines in all cases. While Federated-Clustering

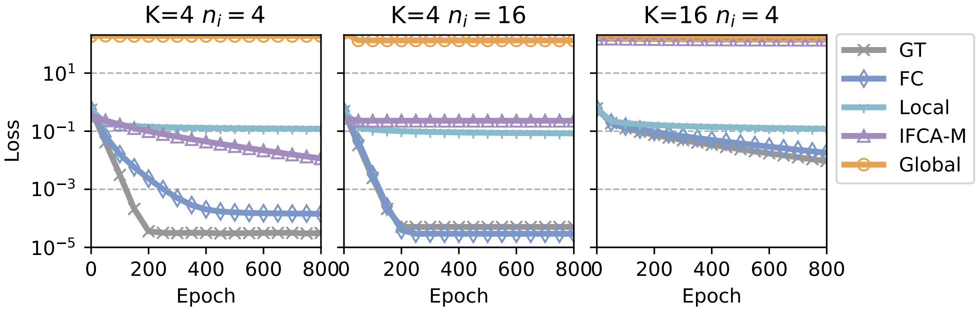


Figure 2.3: The performance of our algorithm vs. baselines on a synthetic dataset. When n_i is small, the ground-truth outperforms our algorithm, but this difference vanishes with increasing n_i. This behavior is consistent with the dependence of the convergence rate on n_i in Theorem 3: increasing n_i improves convergence.

Table 2.1: Comparison of test losses and accuracies for federated personalization algorithms on MNIST. FC outperforms all non-oracle baselines on two learning tasks.

	Rotation		Private label	
	Acc.(%)	Loss	Acc.(%)	Loss
Local	71.3	0.517	75.2	0.489
Global	46.6	0.631	22.2	0.803
Ditto	62.0	0.576	61.7	0.578
IFCA	54.6	0.588	65.4	0.531
KNN	52.1	2.395	63.2	1.411
FC (ours)	**75.4**	**0.475**	**77.0**	**0.468**
GT (oracle)	84.7	0.432	85.1	0.430

is slightly worse than ground truth when $(K, n_i) = (4, 4)$, their performances are almost identical in the middle subplot for $n_i = 16$. This observation is consistent with the n_i-scaling observed in (2.3): as the number of clients-per-cluster increases, convergence improves.

MNIST experiment

In this section, we compare Federated-Clustering to existing federated learning baselines on the MNIST dataset. The dataset is constructed as follows, similar to [10]. The data samples are randomly shuffled and split into $K = 4$ clusters with $n_i = 75$ clients in each cluster. We consider two different tasks: 1) the *rotation* task transforms images in cluster k by $k * 90$ degrees; 2) the *private label* task transforms labels in cluster k with $T_k(y) : y \mapsto (y + k$

mod 10), such that the same image may have different labels from cluster-to-cluster.

Algorithm hyperparameters. For these two experimental tasks, in addition to the baselines from our synthetic experiment, we include the KNN-personalization [25] and Ditto [13] algorithms which both interpolate between a local and global model. The KNN-personalization is a linear combination of a global model, trained with FedAvg [1], and a local model which is the aggregation of nearest-neighbor predictions in the client's local dataset to the global model's prediction. We set the coefficients of this linear combination to be $\lambda_{\text{knn}} = 0.5$ and $\lambda_{\text{knn}} = 0.9$ for the *rotation* and *private label* tasks, respectively. The Ditto objective is a personalized loss with an added regularization term that encourages closeness between the personalized and global models. Since tuning this regularization parameter λ_{ditto} leads to a degenerated "Local" training where $\lambda_{\text{ditto}} = 0$, we fix $\lambda_{\text{ditto}} = 1$ for both *rotation* and *private label* tasks. To reduce the computation cost of our algorithm, in each iteration we randomly divide the N clients into 16 subgroups and apply Federated-Clustering to each subgroup simultaneously. The clipping radius $\tau_{k,l}$ for each cluster k is adaptively chosen to be the 20th-percentile of distances to the cluster-center.

Performance. The experimental results are listed in Table 2.1. Since an image can have different labels across clusters in the *private label* task, a model trained over the pool of all datasets only admits inferior performance. Therefore, distributed training algorithms that maintain a global model, such as FedAvg, Ditto, and KNN, perform poorly compared to training alone. On the other hand, our algorithm Federated-Clustering outperforms standalone training and all personalization baselines. This experiment suggests that our algorithm successfully explored the cluster structure and benefited from collaborative training.

CIFAR experiment

In this section, we evaluate the efficacy of various clustering algorithms on the CIFAR-10 and CIFAR-100 datasets [29].

CIFAR-10

For the CIFAR-10 experiment, we create 4 clusters, each containing 5 clients and transform the labels in each cluster such that different clusters can have different labels for the same image (the *private label* task in Section 2.4). We train a VGG-16 model [31] with batch size 32, learning rate 0.1, and momentum 0.9. The outcomes are presented in Fig. 2.4. The left subplot illustrates that collaborative clustering algorithms designed for global model training (e.g., Ditto, IFCA, Global) yield suboptimal models, as not all participating clients benefit from each other. On the other hand, Local and GroundTruth training are not influenced by the conflicting labels from other clusters so they significantly outperform Ditto and IFCA. Our Federated-Clustering (FC) algorithm also excludes such adversarial influence and, more importantly, outperforms Local training, showing that FC benefits from collaboration.

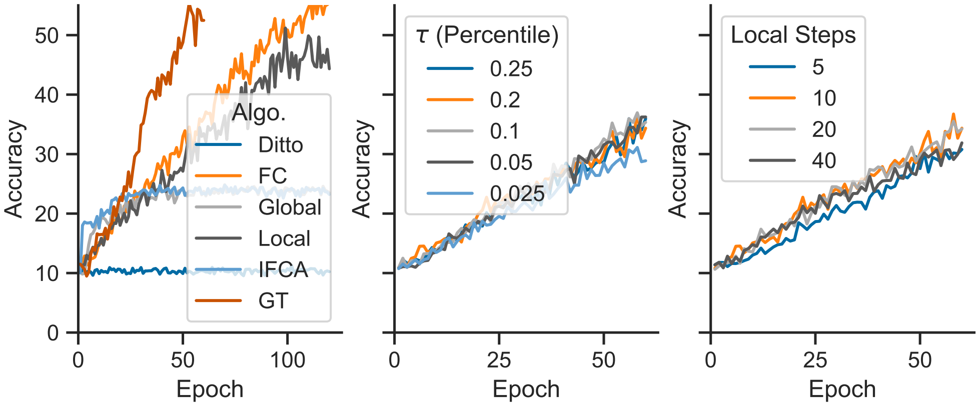


Figure 2.4: Performance of algorithms on CIFAR-10 dataset with the *private labels* task. **Left**: Relative accuracy of clustering algorithms. Algorithms optimized for global model performance, such as Ditto, IFCA, Global (FedAvg), perform poorly on personalization. FC outperforms Local training, showing that it benefits from collaboration between clients, and is competitive with GroundTruth. **Middle**: Impact of thresholding radius τ on accuracy. τ is the percentile of gradients distances from the cluster-center. **Right**: Impact of local gradient steps between two clustering calls. Early on in training, clusters are less identifiable so local optimization helps but these gains lessen later on when gradients from different clusters drift apart and clusters are better defined.

In the middle subplot, we examine the impact on accuracy of varying the thresholding radius τ (i.e. τ is set as the percentile of gradient distances from the cluster-center, so smaller percentile corresponds to smaller τ). Our findings indicate that adopting a more conservative value for τ (lower percentile) does not substantially compromise accuracy.

The right subplot demonstrates the behavior of Federated-Clustering when the clustering oracle is invoked intermittently. The results suggest that increasing the number of local iterations boosts the learning curve early in training when gradients from different clusters are close together and clusters are ill-defined. However, this improvement plateaus when gradients become separated and clusters become well-defined.

CIFAR-100

We consider the CIFAR-100 dataset distributed over 10 clusters so that each cluster contains 10 unique labels. In each cluster, we set 10 clients with IID data. We use a VGG-8 model for training and the same hyperparameters as those in the CIFAR-10 experiment (Section 2.4) and report the results in Fig. 2.5. While clients' data within each cluster share similar features, a small model like VGG-8 cannot sufficiently benefit from intra-cluster collaboration. Therefore the performance of Global training plateaus at a very low level while in contrast Federated-Clustering (FC) still benefits from collaboration and continues to improve over

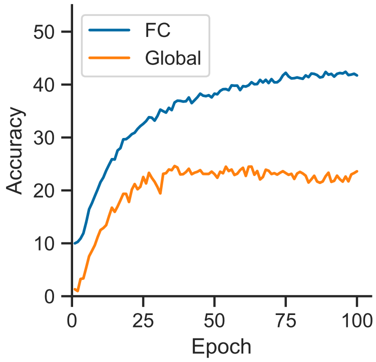


Figure 2.5: Performance of Federated-Clustering (FC) on the CIFAR-100 dataset. The Global (FedAvg) accuracy plateaus while FC continually improves.

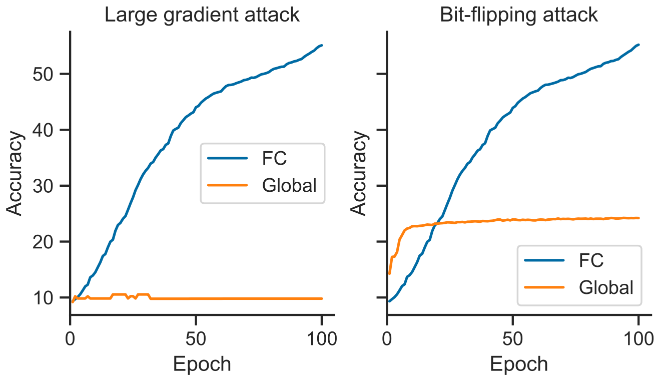


Figure 2.6: Performance of Federated-Clustering (FC) against a large gradient attack (left) and bit-flipping attack (right). FC is robust to these attacks and significantly outperforms Global (FedAvg) performance.

time.

Defense against Byzantine attacks

Byzantine attacks, in which attackers have full knowledge of the system and can deviate from the prescribed algorithm, are prevalent in distributed environments [14]. There are many forms of Byzantine attacks. For example, our *private label* setting in Sections 2.4 and 2.4 corresponds to the *label-flipping attack* in the Byzantine-robustness literature, since a malicious client can try to corrupt the model by assigning the wrong label to an image in training data.

In this section, we investigate two other attacks: Byzantine workers send either very large gradients or gradients with opposite signs. Using the MNIST dataset with the *private label* task, we set 4 clusters with 50 non-malicious clients each (so non-malicious clients from different clusters can have private labels) and add 50 Byzantine works to each cluster. We demonstrate the robustness of Federated-Clustering (FC) in Fig. 2.6. In both cases, Global training suffers from serious model degradation while FC successfully reaches high accuracy under these attacks, demonstrating its robustness.

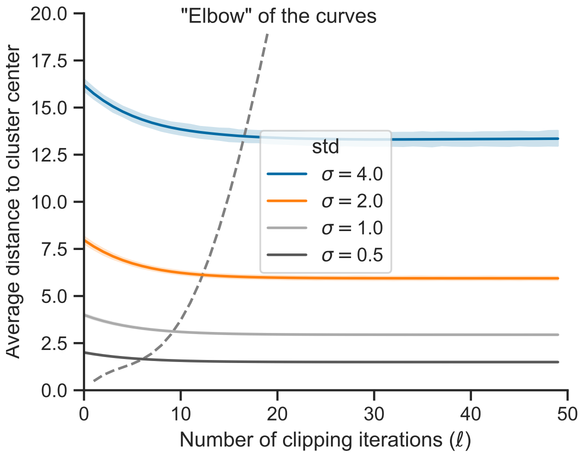


Figure 2.7: The average distance to cluster centers as a function of the number of clipping iterations l. The distance between cluster-centers (Δ) is fixed while inter-cluster standard deviation σ differs.

Empirical Study on Clipping Iterations in Algorithm 3

We employ Algorithm 3 (Threshold-Clustering) on datasets to discern the effectiveness of the clipping iterations in identifying optimal cluster centers. These datasets share the same groundtruth cluster centers, and thus the same Δ, but vary in their inter-cluster standard deviations, with σ values of $0.5, 1, 2, 4$. Each dataset is made up of 90 ten-dimensional samples from 10 clusters, generated using the scikit-learn package [32]. For each iteration l within cluster k, the clipping radius $\tau_{k,l}$ is defined as the 10th percentile of gradients' distances from the cluster-center. We repeat this experimental setup ten times for consistency.

The outcomes, presented in Fig. 2.7, show that the average distances initially decrease rapidly, then steadily approach convergence. To identify the *elbow* of a given curve f, we use the formula $\frac{f(l)-f(l-1)}{f(l-1)-\min_l f(l)}$, where curves post-*elbow* are notably flat. These *elbows* elucidate the correlation between σ and l, indicating that for a fixed dataset (and its corresponding σ), one can pinpoint the minimal iterations l needed for convergence. Notably, this observation appears to align with the $l \gtrsim \log \sigma$ lower bound stated in Theorem 1.

2.5 Conclusion

We develop gradient-based clustering algorithms to achieve personalization in federated learning. Our algorithms have optimal convergence guarantees. They asymptotically match the achievable rates when the true clustering of clients is known, and our analysis holds under light assumptions (e.g., for all smooth convex and non-convex losses). Furthermore, our algorithms are provably robust in the Byzantine setting where some fraction of the clients can arbitrarily corrupt their gradients. Future directions involve developing bespoke analysis for the convex-loss case and developing more communication-efficient versions of our algorithms. Further, our analysis can be used to show that our algorithms are incentive-compatible and lead to *stable coalitions* as in [33]. This would form a strong argument towards encouraging participants in a federated learning system. Investigating such incentives and fairness concerns is another promising future direction.

2.6 Proofs of Theoretical Results

Proof of Theorem 1

First we establish some notation.

Notation.

- $\mathcal{G}_i$ are the good points and $\mathcal{B}_i$ the bad points from point z_i's perspective. Therefore $|\mathcal{G}_i| + |\mathcal{B}_i| = N$.
- k_i denotes the cluster to which client i is assigned at the end of Threshold-Clustering.
- To facilitate the proof, we introduce a variable $c_{k_i,l}^2$ that quantifies the distance from the cluster-center-estimates to the true cluster means at each step of thresholding. Specifically, for client i's cluster k_i at round l of Threshold-Clustering, we set
$$c_{k_i,l}^2 = \mathbb{E}\|v_{k_i,l-1} - \mathbb{E}z_i\|^2.$$
- For client i's cluster k_i at round l of Threshold-Clustering, we use thresholding radius
$$\tau_{k_i,l}^2 \approx c_{k_i,l}^2 + \delta_i \sigma \Delta.$$
- We introduce a variable $y_{j,l}$ to denote the points clipped by Threshold-Clustering:
$$v_{k,l} = \frac{1}{N} \sum_{j\in[N]} \underbrace{z_j \mathbb{1}(\|z_j - v_{k,l-1}\| \leq \tau_{k,l}) + v_{k,l-1}\mathbb{1}(\|z_j - v_{k,l-1}\| > \tau_{k,l})}_{y_{j,l}}.$$

Proof of Theorem 1. We prove the main result with the following sequence of inequalities, and then justify the labeled steps afterwards.

$$\mathbb{E}\|v_{k_i,l} - \mathbb{E}z_i\|^2 = \mathbb{E}\left\|\frac{1}{N}\sum_{j\in[N]} y_{j,l} - \mathbb{E}z_i\right\|^2$$

$$
\begin{aligned}
&= \mathbb{E}\left\|(1-\beta_i)\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l} - \mathbb{E}z_i\right) + \beta_i\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l} - \mathbb{E}z_i\right)\right\|^2 \\
&\overset{(i)}{\leq} (1+\beta_i)(1-\beta_i)^2\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l}\right) - \mathbb{E}z_i\right\|^2 \\
&\qquad + \left(1+\frac{1}{\beta_i}\right)\beta_i^2\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l}\right) - \mathbb{E}x_i\right\|^2 \\
&\lesssim \underbrace{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l}\right) - \mathbb{E}z_i\right\|^2}_{\mathcal{E}_1} + \beta_i \underbrace{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l}\right) - \mathbb{E}z_i\right\|^2}_{\mathcal{E}_2} \\
&\overset{(ii)}{\lesssim} \left((1-\delta_i)c_{k_i,l}^2 + \frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta}\right) + \beta_i(c_{k_i,l}^2 + \delta_i\sigma\Delta) \\
&\overset{(iii)}{\lesssim} (1 - {}^{\delta_i}\!/_2)c_{k_i,l}^2 + \left(\frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \beta_i\sigma\Delta\right) \\
&\overset{(iv)}{\lesssim} (1 - {}^{\delta_i}\!/_2)^l c_{k_i,1}^2 + \left(\frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \beta_i\sigma\Delta\right)\sum_{q=0}^{l-1}(1 - {}^{\delta_i}\!/_2)^q \\
&\overset{(v)}{\lesssim} (1 - {}^{\delta_i}\!/_2)^l \sigma^2 + \left(\frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \beta_i\sigma\Delta\right) \\
&\overset{(vi)}{\lesssim} \frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \beta_i\sigma\Delta.
\end{aligned} \tag{2.9}
$$

Justifications for the labeled steps are:

- (i) Young's inequality: $\|x+y\|^2 \leq (1+\epsilon)\|x\|^2 + (1+{}^1\!/_\epsilon)\|y\|^2$ for any $\epsilon > 0$.
- (ii) We prove this bound in Lemmas 2 and 6. Importantly, it shows that the clustering error is composed of two quantities: $\mathcal{E}_1$, the error contributed by good points from the cluster's perspective, and $\mathcal{E}_2$, the error contributed by the bad points from the cluster's perspective.
- (iii) Assumption that $\beta_i \lesssim \delta_i$
- (iv) Since $\mathbb{E}\|v_{k_i,l} - \mathbb{E}z_i\|^2 = c_{k_i,l+1}^2$, the inequality forms a recursion which we unroll over l steps.
- (v) Assumption that $c_{k_i,1}^2 = \mathbb{E}\|v_{k_i,0} - \mathbb{E}z_i\|^2 \leq \sigma^2$. Also, the partial sum in the second term can be upper-bounded by a large-enough constant.
- (vi) Assumption that $l \geq \max\left\{1, \frac{\log(\sigma/\Delta)}{\log(1-\delta_i/2)}\right\}$

From (2.9), we see that $c_{k_i,l}^2 \lesssim \frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \beta_i \sigma \Delta$. Plugging this into the expression for $\tau_{k_i,l}^2$ gives $\tau_{k_i,l}^2 \approx \frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \delta_i \sigma \Delta \approx \delta_i \sigma \Delta$ for large n_i and Δ. □

Lemma 2 (Clustering Error due to Good Points)**.**

$$\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l}\right) - \mathbb{E}z_i\right\|^2 \lesssim (1-\delta_i)c_{k_i,l}^2 + \frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta}$$

Proof of Lemma 2. We prove the main result with the following sequence of inequalities and justify the labeled steps afterward.

$$\begin{aligned}
\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l}\right) - \mathbb{E}z_i\right\|^2 &= \mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-\mathbb{E}z_j)\right) + \left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\not\sim i}(y_{j,l}-\mathbb{E}z_i)\right)\right\|^2 \\
&\overset{(i)}{\leq} \left(1+\frac{2}{\delta_i}\right)\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-\mathbb{E}z_j)\right\|^2 \\
&\quad + \left(1+\frac{\delta_i}{2}\right)\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\not\sim i}(y_{j,l}-\mathbb{E}z_i)\right\|^2 \\
&\overset{(ii)}{\lesssim} \left(1+\frac{2}{\delta_i}\right)\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(\mathbb{E}y_{j,l}-\mathbb{E}z_j)\right\|^2 \\
&\quad + \left(1+\frac{2}{\delta_i}\right)\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-\mathbb{E}y_{j,l})\right\|^2 \\
&\quad + \left(1+\frac{\delta_i}{2}\right)\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\not\sim i}(y_{j,l}-\mathbb{E}z_i)\right\|^2 \\
&\leq \left(1+\frac{2}{\delta_i}\right)\delta_i^2\|\mathbb{E}_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-z_j)\|^2 \\
&\quad + \left(1+\frac{2}{\delta_i}\right)\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-\mathbb{E}y_{j,l})\right\|^2 \\
&\quad + \left(1+\frac{\delta_i}{2}\right)(1-\delta_i)^2\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l}-\mathbb{E}z_i\|^2 \\
&\lesssim \delta_i \underbrace{\|\mathbb{E}_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-z_j)\|^2}_{\mathcal{T}_i} + \left(1+\frac{2}{\delta_i}\right)\underbrace{\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-\mathbb{E}y_{j,l})\right\|^2}_{\mathcal{T}_2} \\
&\quad + \left(1+\frac{\delta_i}{2}\right)(1-\delta_i)^2\underbrace{\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l}-\mathbb{E}z_i\|^2}_{\mathcal{T}_3}
\end{aligned}$$

$$
\begin{aligned}
&\overset{(iii)}{\lesssim} \delta_i\left(c_{k_i,l}^2 + \frac{(c_{k_i,l}^2+\sigma^2)\sigma}{\delta_i\Delta}\right) + \left(1+\frac{2}{\delta_i}\right)\frac{n_i}{|\mathcal{G}_i|^2}\sigma^2 \\
&\qquad + (1-\delta_i)^2\left(1+\frac{\delta_i}{2}\right)\left(\left(1+\frac{\delta_i}{2}+\frac{\sigma^2}{\delta_i\Delta^2}\right)c_{k_i,l}^2 + \frac{\sigma^3}{\Delta}\right) \\
&\lesssim \left(1-\delta_i+\frac{\sigma}{\Delta}+\frac{\sigma^2}{\delta_i\Delta^2}\right)c_{k_i,l}^2 + \left(\frac{\sigma^2}{n_i}+\frac{\sigma^3}{\Delta}\right) \\
&\overset{(iv)}{\lesssim} (1-\delta_i)c_{k_i,l}^2 + \left(\frac{\sigma^2}{n_i}+\frac{\sigma^3}{\Delta}\right).
\end{aligned}
$$

- (i), (ii) Young's inequality
- (iii) We prove this bound in Lemmas 3, 4, and 5. Importantly, it shows that, from point i's perspective, the error of its cluster-center-estimate is composed of three quantities: $\mathcal{T}_1$, the error introduced by our thresholding procedure on the good points which belong to i's cluster (and therefore ideally are included within the thresholding radius); $\mathcal{T}_2$, which accounts for the variance of the points in i's cluster; and $\mathcal{T}_3$, the error due to the good points which don't belong to i's cluster (and therefore ideally are forced outside the thresholding radius).
- (iv) Assumption that $\Delta \gtrsim \sigma/\delta_i$.

□

Lemma 3 (Bound $\mathcal{T}_1$: Error due to In-Cluster Good Points)**.**

$$\|\mathbb{E}_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-z_j)\|^2 \lesssim c_{k_i,l}^2 + \frac{(c_{k_i,l}^2+\sigma^2)\sigma}{\delta_i\Delta}.$$

Proof of Lemma 3. By definition of $y_{j,l}$,

$$
\begin{aligned}
\mathbb{E}_{j\in\mathcal{G}_i:j\sim i}\|y_{j,l}-z_j\| &= \mathbb{E}[\|v_{k_i,l-1}-z_j\|\mathbb{1}(\|v_{k_i,l-1}-z_j\|>\tau_{k_i,l})] \\
&\le \frac{\mathbb{E}[\|v_{k_i,l-1}-z_j\|^2\mathbb{1}(\|v_{k_i,l-1}-z_j\|>\tau_{k_i,l})]}{\tau_{k_i,l}} \\
&\le \frac{\mathbb{E}\|v_{k_i,l-1}-z_j\|^2}{\tau_{k_i,l}} \\
&\lesssim \frac{\mathbb{E}\|v_{k_i,l-1}-\mathbb{E}z_i\|^2+\mathbb{E}\|\mathbb{E}z_j-z_j\|^2}{\tau_{k_i,l}} \\
&\le \frac{c_{k_i,l}^2+\sigma^2}{\tau_{k_i,l}}
\end{aligned}
$$

Finally, by Jensen's inequality and plugging in the value for $\tau_{k_i,l}$,

$$\|\mathbb{E}(y_{j,l}-z_j)\|^2 \le (\mathbb{E}\|y_{j,l}-z_j\|)^2$$

$$
\begin{aligned}
&\lesssim \frac{(c_{k_i,l}^2+\sigma^2)^2}{\tau_{k_i,l}^2} \\
&\lesssim \frac{c_{k_i,l}^4}{c_{k_i,l}^2} + \frac{c_{k_i,l}^2\sigma^2}{\delta_i\sigma\Delta} + \frac{\sigma^4}{\delta_i\sigma\Delta} \\
&= c_{k_i,l}^2 + \frac{(c_{k_i,l}^2+\sigma^2)\sigma}{\delta_i\Delta}.
\end{aligned}
$$

□

Lemma 4 (Bound $\mathcal{T}_2$: Variance of Clipped Points)**.**

$$
\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-\mathbb{E}y_{j,l})\right\|^2 \le \frac{n_i}{|\mathcal{G}_i|^2}\sigma^2.
$$

Proof of Lemma 4. Note that the elements in the sum $\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-\mathbb{E}y_{j,l})$ are not independent. Therefore, we cannot get rid of the cross terms when expanding the squared-norm. However, if for each round of thresholding we were sample a fresh batch of points to set the new cluster-center estimate, then the terms would be independent. With this resampling strategy, our bounds would only change by a constant factor. Therefore, for ease of analysis, we will assume the terms in the sum are independent. In that case,

$$
\begin{aligned}
\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l}-\mathbb{E}y_{j,l})\right\|^2 &\le \frac{n_i}{|\mathcal{G}_i|^2}\mathbb{E}\|y_{j,l}-\mathbb{E}y_{j,l}\|^2 \\
&\le \frac{n_i}{|\mathcal{G}_i|^2}\mathbb{E}\|z_j-\mathbb{E}z_j\|^2 \\
&\le \frac{n_i}{|\mathcal{G}_i|^2}\sigma^2,
\end{aligned}
$$

where the second-to-last inequality follows from the contractivity of the thresholding procedure. □

Lemma 5 (Bound $\mathcal{T}_3$: Error due to Out-of-Cluster Good Points)**.**

$$
\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l}-\mathbb{E}z_i\|^2 \lesssim \left(1+\frac{\delta_i}{2}+\frac{\sigma^2}{\delta_i\Delta^2}\right)c_{k_i,l}^2+\frac{\sigma^3}{\Delta}.
$$

Proof of Lemma 5. By Young's inequality,

$$
\begin{aligned}
\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l}-\mathbb{E}z_i\|^2 &\le \left(1+\frac{\delta_i}{2}\right)\mathbb{E}\|v_{k_i,l-1}-\mathbb{E}z_i\|^2+\left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l}-v_{k_i,l-1}\|^2 \\
&\le \left(1+\frac{\delta_i}{2}\right)c_{k_i,l}^2+\left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l}-v_{k_i,l-1}\|^2
\end{aligned}
$$

$$\begin{aligned}
&= \left(1+\frac{\delta_i}{2}\right)c_{k_i,l}^2 \\
&\quad + \left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}[\|z_j - v_{k_i,l-1}\|^2\mathbb{1}\{\|z_j - v_{k_i,l-1}\| \le \tau_{k_i,l}\}] \\
&\le \left(1+\frac{\delta_i}{2}\right)c_{k_i,l}^2 + \left(1+\frac{2}{\delta_i}\right)\tau_{k_i,l}^2\mathbb{P}_{j\in\mathcal{G}_i:j\not\sim i}(\|z_j - v_{k_i,l-1}\| \le \tau_{k_i,l}).
\end{aligned}$$

We now have to bound the probability in the expression above. Note that if $\|v_{k_i,l-1} - z_j\| \le \tau_{k_i,l}$, then

$$\begin{aligned}
\|\mathbb{E}z_j - \mathbb{E}z_i\|^2 &\lesssim \|z_j - \mathbb{E}z_j\|^2 + \|z_j - \mathbb{E}v_{k_i,l-1}\|^2 + \|\mathbb{E}v_{k_i,l-1} - \mathbb{E}z_i\|^2 \\
&\lesssim \|z_j - \mathbb{E}z_j\|^2 + \|z_j - \mathbb{E}z_j\|^2 + \|\mathbb{E}z_j - \mathbb{E}v_{k_i,l-1}\|^2 \\
&\quad + \mathbb{E}\|v_{k_i,l-1} - \mathbb{E}z_i\|^2 + \mathbb{E}\|z_i - \mathbb{E}z_i\|^2 \\
&\lesssim \|z_j - \mathbb{E}z_j\|^2 + \tau_{k_i,l}^2 + c_{k_i,l}^2 + \sigma^2.
\end{aligned}$$

By Assumption 2, this implies that

$$\Delta^2 \lesssim \|z_j - \mathbb{E}z_j\|^2 + \tau_{k_i,l}^2 + c_{k_i,l}^2 + \sigma^2$$

which means that

$$\|z_j - \mathbb{E}z_j\|^2 \gtrsim \Delta^2 - (\tau_{k_i,l}^2 + c_{k_i,l}^2 + \sigma^2).$$

By Markov's inequality,

$$\mathbb{P}(\|z_j - \mathbb{E}z_j\|^2 \gtrsim \Delta^2 - (\tau_{k_i,l}^2 + c_{k_i,l}^2 + \sigma^2)) \le \frac{\sigma^2}{\Delta^2 - (\tau_{k_i,l}^2 + c_{k_i,l}^2 + \sigma^2)} \lesssim \frac{\sigma^2}{\Delta^2}$$

as long as

$$\Delta^2 \gtrsim \tau_{k_i,l}^2 + c_{k_i,l}^2 + \sigma^2,$$

which holds due to the constraint on Δ in the theorem statement. Therefore

$$\begin{aligned}
\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l} - \mathbb{E}z_i\|^2 &\lesssim \left(1+\frac{\delta_i}{2}\right)c_{k_i,l}^2 + \left(1+\frac{2}{\delta_i}\right)\frac{(c_{k_i,l}^2 + \delta_i\sigma\Delta)\sigma^2}{\Delta^2} \\
&\le \left(1+\frac{\delta_i}{2}+\frac{\sigma^2}{\delta_i\Delta^2}\right)c_{k_i,l}^2 + \frac{\sigma^3}{\Delta}.
\end{aligned}$$

□

Lemma 6 (Clustering Error due to Bad Points)**.**

$$\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l}\right) - \mathbb{E}z_i\right\|^2 \lesssim c_{k_i,l}^2 + \delta_i\sigma\Delta$$

Proof of Lemma 6.

$$\begin{aligned}\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l}\right) - \mathbb{E}z_i\right\|^2 &\leq \mathbb{E}_{j\in\mathcal{B}_i}\|y_{j,l} - \mathbb{E}z_i\|^2 \\ &\lesssim \mathbb{E}_{j\in\mathcal{B}_i}\|y_{j,l} - v_{k_i,l-1}\|^2 + \mathbb{E}\|v_{k_i,l-1} - \mathbb{E}z_i\|^2 \\ &\lesssim c_{k_i,l}^2 + \delta_i\sigma\Delta.\end{aligned}$$

The last inequality follows from the intuition that bad points will position themselves at the edge of the thresholding ball, a distance $\tau_{k_i,l}$ away from the current center-estimate $v_{k_i,l-1}$. Therefore we cannot do better than upper-bounding $\mathbb{E}_{j\in\mathcal{B}_i}\|y_{j,l} - v_{k_i,l-1}\|^2$ by $\tau_{k_i,l}^2 \approx c_{k_i,l}^2 + \delta_i\sigma\Delta$, the squared-radius of the ball. □

Proof of Theorem 2

Proof of Theorem 2. Let

$$\mathcal{D}_1 = \begin{cases}\delta & \text{w.p. } p\\ 0 & \text{w.p. } 1-p\end{cases}$$

and

$$\mathcal{D}_2 = \begin{cases}\delta & \text{w.p. } 1-p\\ 0 & \text{w.p. } p\end{cases}$$

and define the mixture $\mathcal{M} = \frac{1}{2}\mathcal{D}_1 + \frac{1}{2}\mathcal{D}_2$. Also consider the mixture $\tilde{\mathcal{M}} = \frac{1}{2}\tilde{\mathcal{D}}_1 + \frac{1}{2}\tilde{\mathcal{D}}_2$, where $\tilde{\mathcal{D}}_1 = 0$ and $\tilde{\mathcal{D}}_2 = \delta$. It is impossible to distinguish whether a sample comes from $\mathcal{M}$ or $\tilde{\mathcal{M}}$. Therefore, if you at least know a sample came from either $\mathcal{M}$ or $\tilde{\mathcal{M}}$ but not which one, the best you can do is to estimate μ_1 with $\hat{\mu}_1 = \frac{\delta p}{2}$, half-way between the mean of $\mathcal{D}_1$, which is δp, and the mean of $\tilde{\mathcal{D}}_1$, which is 0. In this case

$$\mathbb{E}\|\hat{\mu}_1 - \mu_1\|^2 = \frac{\delta^2 p^2}{4}.$$

If $p \leq \frac{1}{2}$, then

$$\Delta = (1-p)\delta - p\delta = (1-2p)\delta. \tag{2.10}$$

Also,

$$\sigma^2 = \delta^2 p(1-p). \tag{2.11}$$

Equating δ^2 in (2.10) and (2.11),

$$\frac{\Delta^2}{(1-2p)^2} = \frac{\sigma^2}{p(1-p)},$$

which can be rearranged to

$$(4\sigma^2 + \Delta^2)p^2 - (4\sigma^2 + \Delta^2)p + \sigma^2 = 0.$$

Solving for p,

$$p = \frac{1}{2} - \frac{\Delta}{2\sqrt{4\sigma^2 + \Delta^2}}. \tag{2.12}$$

Note that,

$$\begin{aligned} \frac{\delta^2 p^2}{4} &= \frac{\sigma^2 p^2}{4p(1-p)} \\ &= \frac{\sigma^2 p}{4(1-p)}. \end{aligned} \tag{2.13}$$

Plugging the expression for p from (2.12) into (2.13), we can see that

$$\frac{\delta^2 p^2}{4} = \frac{\sigma^2}{4}\left(\frac{\sqrt{4\sigma^2+\Delta^2} - \Delta}{\Delta}\right) = \frac{\sigma^2}{4}\left(\sqrt{1 + \frac{4\sigma^2}{\Delta^2}} - 1\right) \geq \frac{\sigma^2}{4}\left(\frac{2\sigma^2}{\Delta^2} - \frac{2\sigma^4}{\Delta^4}\right).$$

The last step used an immediately verifiable inequality that $\sqrt{1+x} \geq 1 + \frac{x}{2} - \frac{x^2}{8}$ for all $x \in [0, 8]$. Finally, we can choose $\Delta^2 \geq 2\sigma^2$ to give the result that

$$\mathbb{E}\|\hat{\mu}_1 - \mu_1\|^2 \geq \frac{\delta^2 p^2}{4} \geq \frac{\sigma^4}{4\Delta^2}\,.$$

Finally, suppose that there is only a single cluster with $K = 1$. Then, given n stochastic samples. standard information theoretic lower bounds show that we will have an error at least

$$\mathbb{E}\|\hat{\mu}_1 - \mu_1\|^2 \geq \frac{\sigma^2}{4n}\,.$$

Combining these two lower bounds yields the proof of the theorem. □

Proof of Lemma 1

Proof. For h an L-smooth function and g a μ-strongly-convex function with shared optimum x^*, the following inequality holds for all x:

$$\|\nabla h(x)\|^2 \leq \left(\frac{L}{\mu}\right)^2 \|\nabla g(x)\|^2. \tag{2.14}$$

To see this, note that by L-smoothness of h

$$\|\nabla h(x)\|^2 = \|\nabla h(x) - \nabla h(x^*)\|^2 \leq L^2\|x - x^*\|^2. \tag{2.15}$$

By μ-strong-convexity of g and Cauchy-Schwarz inequality,

$$\mu\|x - x^*\|^2 \leq \langle \nabla g(x) - \nabla g(x^*), x - x^* \rangle \leq \|\nabla g(x)\|\|x - x^*\|. \tag{2.16}$$

Rearranging terms in (2.16), squaring both sides, and combining it with (2.15) gives (2.14).

We can now apply (2.14) to show that Assumptions 4 and 5 hold.

For Assumption 4, let $h(x) = f_i(x) - \nabla \bar{f}_i(x)$ and $g(x) = \nabla \bar{f}_i(x)$. Thus h and g have the same optimum. Since the average of μ-strongly-convex functions is μ-strongly-convex, g is μ-strongly-convex. By L-smoothness of f_i,

$$\begin{aligned}\|(\nabla f_i(x) - \nabla \bar{f}_i(x)) - (\nabla f_i(y) - \nabla \bar{f}_i(y))\| &\leq \|\nabla f_i(x) - \nabla f_i(y)\| + \|\nabla \bar{f}_i(x)) - \nabla \bar{f}_i(y)\| \\ &\leq 2L\|x - y\|,\end{aligned}$$

showing that h is $2L$-smooth. Therefore, by (2.14)

$$\|\nabla f_i(x) - \nabla \bar{f}_i(x)\|^2 \leq \left(\frac{2L}{\mu}\right)^2 \|\nabla \bar{f}_i(x)\|^2,$$

which shows that Assumption 4 is satisfied with $A = {}^{2L}/_{\mu}$.

For Assumption 5, let x_i^* be client i's optimum (equivalently the optimum of all clients in client i's cluster).

$$\begin{aligned}\|\nabla f_i(x) - \nabla f_j(x)\|^2 &= \|\nabla f_i(x) - (\nabla f_j(x) - \nabla f_j(x_i^*)) - (\nabla f_j(x_i^*) - \nabla f_i(x_i^*)))\|^2 \\ &\overset{(i)}{\geq} \frac{1}{2}\|\nabla f_j(x_i^*) - \nabla f_i(x_i^*)\|^2 - \|\nabla f_i(x) - (\nabla f_j(x) - \nabla f_j(x_i^*))\|^2 \\ &\overset{(ii)}{\geq} \frac{1}{2}\|\nabla f_j(x_i^*) - \nabla f_i(x_i^*)\|^2 - 2\|\nabla f_i(x)\|^2 - 2\|\nabla f_j(x) - \nabla f_j(x_i^*)\|^2 \\ &\overset{(iii)}{\geq} \frac{1}{2}\|\nabla f_j(x_i^*) - \nabla f_i(x_i^*)\|^2 - 2\|\nabla f_i(x)\|^2 - 2({}^{L}/_{\mu})^2\|\nabla f_i(x)\|^2 \\ &= \frac{1}{2}\|\nabla f_j(x_i^*) - \nabla f_i(x_i^*)\|^2 - 2(1 + ({}^{L}/_{\mu})^2)\|\nabla f_i(x)\|^2 \\ &= \frac{1}{2}\|\nabla f_j(x_i^*)\|^2 - 2(1 + ({}^{L}/_{\mu})^2)\|\nabla f_i(x)\|^2, \qquad (2.17)\end{aligned}$$

where justification for the steps are:

- (i) For all a, b, it holds that $(a-b)^2 \geq \frac{1}{2}b^2 - a^2$, since this inequality can be rearranged to state $({}^{b}/_{\sqrt{2}} - \sqrt{2}a)^2 \geq 0$.
- (ii) Young's inequality.
- (iii) Set $h(x) = f_j(x) - \langle f_j(x_i^*), x\rangle$ and $g(x) = f_i(x)$. Then $\nabla h(x) = \nabla f_j(x) - \nabla f_j(x_i^*)$, from which we see that h and g have the same optimum x_i^* and h is L-smooth (since $\|\nabla h(x) - \nabla h(y)\| = \|(\nabla f_j(x) - \nabla f_j(x_i^*)) - (\nabla f_j(y) - \nabla f_j(x_i^*))\| = L\|x - y\|$). Applying (2.14) gives the desired result.

Therefore (2.17) shows that Assumption 5 is satisfied with $\Delta = \frac{1}{\sqrt{2}} \max_{j \not\sim i} \|\nabla f_j(x_i^*)\|$ and $D = \sqrt{2(1 + ({}^{L}/_{\mu})^2)}$. □

Proof of Theorem 3

First we establish some notation.

Notation.

- $\mathcal{G}_i$ are the good clients and $\mathcal{B}_i$ the bad clients from client i's perspective. Therefore $|\mathcal{G}_i| + |\mathcal{B}_i| = N$.
- $\mathbb{E}_x$ denotes conditional expectation given the parameter, e.g. $\mathbb{E}_x g(x) = \mathbb{E}[g(x)|x]$. $\mathbb{E}$ denotes expectation over all randomness.
- $\overline{X_t} \triangleq \frac{1}{T}\sum_{t=1}^{T} X_t$ for a general variable X_t indexed by t.
- We use $f_i(x)$ to denote average loss on a general batch B of samples. That is, if $f_i(x;b)$ is the loss on a single sample b, we define $f_i(x) = \frac{1}{|B|}\sum_{b\in B} f_i(x;b)$.
- $\bar{f}_i(x) \triangleq \frac{1}{n_i}\sum_{j\in\mathcal{G}_i : j\sim i} f_j(x)$
- We introduce a variable ρ^2 to bound the variance of the gradients

$$\mathbb{E}\|g_i(x) - \mathbb{E}_x g_i(x)\|^2 \leq \rho^2,$$

  and show in Lemma 7 how this can be written in terms of the variance of the gradients computed over a batch size of 1.
- l_t is the number of rounds that Threshold-Clustering is run in round t of Federated-Clustering.
- k_i denotes the cluster to which client i is assigned.
- $v_{i,l,t}$ denotes the gradient update for client i in round t of Federated-Clustering and round l of Threshold-Clustering. That is, $v_{i,l_t,t}$ corresponds to the quantity returned in Step 6 of Algorithm 2.
- To facilitate the proof, we introduce a variable $c_{k_i,l,t}$ that quantifies the distance from the cluster-center-estimates to the true cluster means. Specifically, for client i's cluster k_i at round t of Federated-Clustering and round l of Threshold-Clustering we set

$$c^2_{k_i,l,t} = \mathbb{E}\|v_{i,l-1,t} - \mathbb{E}_x \bar{g}_i(x_{i,t-1})\|^2.$$

- For client i's cluster k_i at round t of Federated-Clustering and round l of Threshold-Clustering, we use thresholding radius

$$\tau^2_{k_i,l,t} \approx c^2_{k_i,l,t} + A^4\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2 + \delta_i\rho\Delta.$$

- Finally, we introduce a variable $y_{j,l,t}$ to denote the points clipped by Threshold-Clustering:

$$v_{i,l,t} = \frac{1}{N}\sum_{j\in[N]} \underbrace{\mathbb{1}(\|g_j(x_{i,t-1}) - v_{i,l-1,t}\| \leq \tau_{k_i^t,l}) + v_{i,l-1,t}\mathbb{1}(\|g_j(x_{i,t-1}) - v_{i,l-1,t}\| > \tau_{k_i^t,l})}_{y_{j,l,t}}.$$

Proof of Theorem 3. In this proof, our goal is to bound $\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2}$ for each client i, thus showing convergence. Recall that our thresholding procedure clusters the gradients of clients at each round and estimates the center of each cluster. These estimates are then used to update the parameters of the clusters. Therefore, we expect $\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2}$ to be bounded in terms of the error of this estimation procedure. The following sequence of inequalities shows this.

By L-smoothness of f_i and setting $\eta \leq {}^1/_L$,

$$\begin{aligned} f_i(x_{i,t}) &\leq f_i(x_{i,t-1}) + \langle \nabla f_i(x_{i,t-1}), x_{i,t} - x_{i,t-1}\rangle + \frac{L}{2}\|x_{i,t} - x_{i,t-1}\|^2 \\ &= f_i(x_{i,t-1}) - \eta\langle \nabla f_i(x_{i,t-1}), v_{i,l_t,t}\rangle + \frac{L\eta^2}{2}\|v_{i,l_t,t}\|^2 \\ &= f_i(x_{i,t-1}) + \frac{\eta}{2}\|v_{i,l_t,t} - \nabla f_i(x_{i,t-1})\|^2 - \frac{\eta}{2}\|\nabla f_i(x_{i,t-1})\|^2 - \frac{\eta}{2}(1 - L\eta)\|v_{i,l_t,t}\|^2 \\ &\leq f_i(x_{i,t-1}) + \eta\|v_{i,l_t,t} - \nabla \bar{f}_i(x_{i,t-1})\|^2 + \eta A^2\|\nabla \bar{f}_i(x_{i,t-1})\|^2 \\ &\qquad - \frac{\eta}{2}\|\nabla f_i(x_{i,t-1})\|^2 - \frac{\eta}{2}(1 - L\eta)\|v_{i,l_t,t}\|^2. \end{aligned} \tag{2.18}$$

Recall that $v_{i,l_t,t}$ is client i's cluster-center estimate at round t of optimization, so the second term on the right side of (2.18) is the error due to the clustering procedure. In Lemma 8, we show that, in expectation, this error is bounded by

$$\delta_i\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right).$$

Therefore, subtracting f_i^* from both sides, summing (2.18) over t, dividing by T, taking expectations, applying Lemma 8 to (2.18), and applying the constraint on Δ from the theorem statement, we have

$$\begin{aligned} \eta\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} \lesssim \frac{\mathbb{E}(f_i(x_{i,0}) - f_i^*)}{T} + \eta A^2\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} \\ + \eta\left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right). \end{aligned} \tag{2.19}$$

The third term on the right side of (2.18) reflects the fact that clients in the same cluster may have different loss objectives. Far from their optima, these loss objectives may look very different and therefore be hard to cluster together.

In order to bound this term, we use a similar argument as above. By L-smoothness of f_i's and setting $\eta \leq {}^1\!/\!_L$,

$$\begin{aligned}\bar{f}_i(x_{i,t}) &\leq \bar{f}_i(x_{i,t-1}) + \langle \nabla \bar{f}_i(x_{i,t-1}), x_{i,t} - x_{i,t-1}\rangle + \frac{L}{2}\|x_{i,t} - x_{i,t-1}\|^2 \\ &= \bar{f}_i(x_{i,t-1}) - \eta\langle \nabla \bar{f}_i(x_{i,t-1}), v_{i,l_t,t}\rangle + \frac{L\eta^2}{2}\|v_{i,l_t,t}\|^2 \\ &= \bar{f}_i(x_{i,t-1}) + \frac{\eta}{2}\|v_{i,l_t,t} - \nabla \bar{f}_i(x_{i,t-1})\|^2 - \frac{\eta}{2}\|\nabla \bar{f}_i(x_{i,t-1})\|^2 - \frac{\eta}{2}(1 - L\eta)\|v_{i,l_t,t}\|^2.\end{aligned}$$

Subtracting $\bar{f}_i^*$ from both sides, summing over t, dividing by T, taking expectations, and applying Lemma 8,

$$\begin{aligned}\eta\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} \lesssim{}& \frac{\mathbb{E}(\bar{f}_i(x_{i,0}) - \bar{f}_i^*)}{T} + \frac{\eta\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} \\ &+ \eta\left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right).\end{aligned} \tag{2.20}$$

Combining (2.19) and (2.20), and applying the constraint on Δ from the theorem statement, we have that

$$\begin{aligned}\eta\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} \lesssim{}& \frac{\mathbb{E}(f_i(x_{i,0}) - f_i^*) + A^2\mathbb{E}(\bar{f}_i(x_0) - \bar{f}_i^*)}{T} \\ &+ \eta\max(1, A^2)\left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right).\end{aligned} \tag{2.21}$$

Dividing both sides of (2.21) by $\eta = {}^1\!/\!_L$ and noting that $\rho^2 = {}^{\sigma^2}\!/\!_{|B|}$ from Lemma 7,

$$\begin{aligned}\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} &\lesssim \frac{\mathbb{E}(f_i(x_{i,0}) - f_i^*) + A^2\mathbb{E}(\bar{f}_i(x_{i,0}) - \bar{f}_i^*)}{\eta T} \\ &\qquad + \max(1, A^2)\left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right) \qquad (2.22) \\ &\leq \frac{L(\mathbb{E}(f_i(x_{i,0}) - f_i^*) + A^2\mathbb{E}(\bar{f}_i(x_{i,0}) - \bar{f}_i^*))}{T} \\ &\qquad + \frac{\max(1, A^2)({}^{\sigma^2}\!/\!_{n_i} + {}^{\sigma^3}\!/\!_{\Delta} + \beta_i\sigma\Delta)}{\sqrt{|B|}} \\ &\lesssim \sqrt{\frac{\max(1, A^2)({}^{\sigma^2}\!/\!_{n_i} + {}^{\sigma^3}\!/\!_{\Delta} + \beta_i\sigma\Delta)}{T}},\end{aligned}$$

where the last inequality follows from setting

$$|B| \gtrsim \max(1, A^2)\left(\frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \beta_i\sigma\Delta\right)T.$$

Since $T = {}^{m_i}/_{|B|}$, this is equivalent to setting

$$|B| \gtrsim \sqrt{\max(1, A^2)\left(\frac{\sigma^2}{n_i} + \frac{\sigma^3}{\Delta} + \beta_i \sigma \Delta\right) m_i}.$$

□

Lemma 7 (Variance reduction using batches)**.** *If, for a single sample b,*

$$\mathbb{E}_x \|g_i(x; b) - \mathbb{E}_x g_i(x; b)\|^2 \leq \sigma^2,$$

then for a batch B of samples,

$$\mathbb{E}_x \|g_i(x) - \mathbb{E}_x g_i(x)\|^2 \leq \frac{\sigma^2}{|B|}.$$

Proof of Lemma 7. Due to the independence and unbiasedness of stochastic gradients,

$$\begin{aligned}\mathbb{E}_x \|g_i(x) - \mathbb{E}_x g_i(x)\|^2 &= \frac{1}{|B|^2} \sum_{b \in B} \mathbb{E}_x \|g_i(x; b) - \mathbb{E}_x g_i(x; b)\|^2 \\ &\leq \frac{\sigma^2}{|B|}.\end{aligned}$$

□

Lemma 8 (Bound on Clustering Error)**.**

$$\overline{\mathbb{E}\|v_{i,l_t,t} - \nabla \bar{f}_i(x_{i,t-1})\|^2} \lesssim \delta_i \overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta} D^2 \overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i \rho \Delta\right)$$

Proof of Lemma 8. We prove the main result with the following sequence of inequalities and justify the labeled steps afterwards.

$$\begin{aligned}&\overline{\mathbb{E}\|v_{i,l_t,t} - \nabla \bar{f}_i(x_{i,t-1})\|^2} \\ &= \overline{\mathbb{E}\|v_{i,l_t,t} - \mathbb{E}_x \bar{g}_i(x_{i,t-1})\|^2} \\ &= \overline{\mathbb{E}\left\|\frac{1}{N} \sum_{j \in [N]} y_{j,l_t,t} - \mathbb{E}_x \bar{g}_i(x_{i,t-1})\right\|^2} \\ &= \overline{\mathbb{E}\left\|(1 - \beta_i)\left(\frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i} y_{j,l_t,t} - \mathbb{E}_x \bar{g}_i(x_{i,t-1})\right) + \beta_i \left(\frac{1}{|\mathcal{B}_i|} \sum_{j \in \mathcal{B}_i} y_{j,t,l} - \mathbb{E}_x \bar{g}_i(x_{i,t-1})\right)\right\|^2} \\ &\overset{(i)}{\leq} (1 + \beta_i)(1 - \beta_i)^2 \overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i} y_{j,l_t,t}\right) - \mathbb{E}_x \bar{g}_i(x_{i,t-1})\right\|^2}\end{aligned}$$

$$
\begin{aligned}
&+\left(1+\frac{1}{\beta_i}\right)\beta_i^2\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l_t,t}\right)-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\right\|^2} \\
&\leq \underbrace{\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t}\right)-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\right\|^2}}_{\mathcal{E}_1} + \beta_i \underbrace{\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l_t,t}\right)-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\right\|^2}}_{\mathcal{E}_2} \\
&\overset{(ii)}{\lesssim} (1-\delta_i+\beta_i)\overline{c^2_{k_i,l_t,t}} + (\delta_i+\beta_i A^4)\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} \\
&\quad + \left(\frac{\rho^2}{n_i}+\frac{\rho^3}{\Delta}+\beta_i\rho\Delta\right) \\
&\overset{(iii)}{\lesssim} (1-{}^{\delta_i}\!/_2)\overline{c^2_{k_i,l_t,t}} + \delta_i\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \left(\frac{\rho^2}{n_i}+\frac{\rho^3}{\Delta}+\beta_i\rho\Delta\right) \\
&\overset{(iv)}{\lesssim} \overline{(1-{}^{\delta_i}\!/_2)^{l_t}c^2_{k_i,1,t}} + \delta_i\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \left(\frac{\rho^2}{n_i}+\frac{\rho^3}{\Delta}+\beta_i\rho\Delta\right) \\
&\overset{(v)}{\leq} \frac{\rho}{\Delta}\overline{c^2_{k_i,1,t}} + \delta_i\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \left(\frac{\rho^2}{n_i}+\frac{\rho^3}{\Delta}+\beta_i\rho\Delta\right). \qquad (2.23)
\end{aligned}
$$

Justifications for the labeled steps are:

- (i) Young's inequality: $\|x+y\|^2 \leq (1+\epsilon)x^2 + (1+{}^1\!/_\epsilon)y^2$ for any $\epsilon > 0$.
- (ii) We prove this bound in Lemmas 9 and 13. Importantly, it shows that the clustering error is composed of two quantities: $\mathcal{E}_1$, the error contributed by good points from the cluster's perspective, and $\mathcal{E}_2$, the error contributed by the bad points from the cluster's perspective.
- (iii) Assumption that $\beta_i \lesssim \min(\delta_i, {}^{\delta_i}\!/_{A^4})$
- (iv) Since $\mathbb{E}\|v_{i,l_t,t} - \nabla \bar{f}_i(x_{i,t-1})\|^2 = c^2_{k_i,l_t+1,t}$, the inequality forms a recursion which we unroll over l_t steps.
- (v) Assumption that $l_t \geq \max(1, \frac{\log({}^\rho\!/_\Delta)}{\log(1-{}^{\delta_i}\!/_2)})$

Finally, we note that

$$
\begin{aligned}
c^2_{k_i,1,t} &= \mathbb{E}\|g_i(x_{i,t-1}) - \mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2 \\
&\lesssim \mathbb{E}\|g_i(x_{i,t-1}) - \mathbb{E}_x g_i(x_{i,t-1})\|^2 + \mathbb{E}\|\mathbb{E}_x g_i(x_{i,t-1}) - \mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2 \\
&\leq \rho^2 + A^2\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2.
\end{aligned}
$$

Applying this bound to (2.23), and applying the bound on Δ from the theorem statement, we have

$$
\overline{\mathbb{E}\|v_{i,l_t,t} - \nabla \bar{f}_i(x_{i,t-1})\|^2}
$$

$$
\begin{aligned}
&\lesssim \left(\delta_i + \frac{\rho A^2}{\Delta}\right)\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right) \\
&\lesssim \delta_i\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right).
\end{aligned}
$$

□

Lemma 9 (Clustering Error due to Good Points)**.**

$$
\begin{aligned}
&\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t}\right) - \mathbb{E}_x\bar{g}_i(x_{i,t-1})\right\|^2} \\
&\quad \lesssim (1-\delta_i)\overline{c^2_{k_i,l_t,t}} + \delta_i\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta}\right)
\end{aligned}
$$

Proof of Lemma 9. We prove the main result in the sequence of inequalities below and then justify the labeled steps.

$$
\begin{aligned}
&\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t}\right) - \mathbb{E}_x\bar{g}_i(x_{i,t-1})\right\|^2} \\
&= \overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t}\right) - \frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}\mathbb{E}_x g_j(x_{i,t-1}) - \left(\frac{1}{n_i} - \frac{1}{|\mathcal{G}_i|}\right)\sum_{j\in\mathcal{G}_i:j\sim i}\mathbb{E}_x g_j(x_{i,t-1})\right\|^2} \\
&= \overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}_x g_j(x_{i,t-1}))\right) + \left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\not\sim i}(y_{j,l_t,t} - \mathbb{E}_x\bar{g}_i(x_{i,t-1}))\right)\right\|^2} \\
&\overset{(i)}{\leq} \left(1+\frac{2}{\delta_i}\right)\overline{\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}_x g_j(x_{i,t-1}))\right\|^2} \\
&\quad + \left(1+\frac{\delta_i}{2}\right)\overline{\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\not\sim i}(y_{j,l_t,t} - \mathbb{E}_x\bar{g}_i(x_{i,t-1}))\right\|^2} \\
&\overset{(ii)}{\lesssim} \left(1+\frac{2}{\delta_i}\right)\overline{\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(\mathbb{E}y_{j,l_t,t} - \mathbb{E}_x g_j(x_{i,t-1}))\right\|^2} \\
&\quad + \left(1+\frac{2}{\delta_i}\right)\overline{\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t})\right\|^2} \\
&\quad + \left(1+\frac{\delta_i}{2}\right)\overline{\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\not\sim i}(y_{j,l_t,t} - \mathbb{E}_x\bar{g}_i(x_{i,t-1}))\right\|^2} \\
&\overset{(iii)}{\lesssim} \left(1+\frac{2}{\delta_i}\right)\overline{\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}\mathbb{E}(y_{j,l_t,t} - g_j(x_{i,t-1}))\right\|^2}
\end{aligned}
$$

$$
\begin{aligned}
&+ \left(1 + \frac{2}{\delta_i}\right) \overline{\mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (\mathbb{E}_x g_j(x_{i,t-1}) - \mathbb{E} g_j(x_{i,t-1})) \right\|^2} \\
&+ \left(1 + \frac{2}{\delta_i}\right) \overline{\mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (y_{j,l_t,t} - \mathbb{E} y_{j,l_t,t}) \right\|^2} \\
&+ \left(1 + \frac{\delta_i}{2}\right) \overline{\mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \not\sim i} (y_{j,l_t,t} - \mathbb{E}_x \bar{g}_i(x_{i,t-1})) \right\|^2} \\
\overset{(iv)}{\lesssim}\ & \left(1 + \frac{2}{\delta_i}\right) \delta_i^2 \underbrace{\overline{\mathbb{E}_{j \in \mathcal{G}_i : j \sim i} \| \mathbb{E}(y_{j,l_t,t} - g_j(x_{i,t-1})) \|^2}}_{\mathcal{T}_1} + \left(1 + \frac{2}{\delta_i}\right) \frac{n_i}{|\mathcal{G}_i|^2} \rho^2 \\
&+ \left(1 + \frac{2}{\delta_i}\right) \underbrace{\overline{\mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (y_{j,l_t,t} - \mathbb{E} y_{j,l_t,t}) \right\|^2}}_{\mathcal{T}_2} \\
&+ \left(1 + \frac{\delta_i}{2}\right) (1 - \delta_i)^2 \underbrace{\overline{\mathbb{E}_{j \in \mathcal{G}_i : j \not\sim i} \| y_{j,l_t,t} - \mathbb{E}_x \bar{g}_i(x_{i,t-1}) \|^2}}_{\mathcal{T}_3} \\
\overset{(v)}{\lesssim}\ & \delta_i \left(\overline{c^2_{k_i,l_t,t}} + \overline{\mathbb{E} \| \nabla \bar{f}_i(x_{i,t-1}) \|^2} + \frac{\rho^3}{\delta_i \Delta} \right) + \left(1 + \frac{2}{\delta_i}\right) \frac{n_i}{|\mathcal{G}_i|^2} \rho^2 \\
&+ \left(1 + \frac{\delta_i}{2}\right) (1 - \delta_i)^2 \left(\left(1 + \frac{\delta_i}{2}\right) \overline{c^2_{k_i,l_t,t}} + \delta_i \overline{\mathbb{E} \| \nabla \bar{f}_i(x_{i,t-1}) \|^2} \right. \\
&\left. + \frac{\rho}{\Delta} D^2 \overline{\mathbb{E} \| \nabla f_i(x_{i,t-1}) \|^2} + \frac{\rho^3}{\Delta} \right) \\
\overset{(vi)}{\lesssim}\ & \delta_i \left(\overline{c^2_{k_i,l_t,t}} + \overline{\mathbb{E} \| \nabla \bar{f}_i(x_{i,t-1}) \|^2} + \frac{\rho^3}{\delta_i \Delta} \right) + \frac{\rho^2}{n_i} \\
&+ (1 - \delta_i) \left(\overline{c^2_{k_i,l_t,t}} + \delta_i \overline{\mathbb{E} \| \nabla \bar{f}_i(x_{i,t-1}) \|^2} + \frac{\rho}{\Delta} D^2 \overline{\mathbb{E} \| \nabla f_i(x_{i,t-1}) \|^2} + \frac{\rho^3}{\Delta} \right) \\
\lesssim\ & (1 - \delta_i) \overline{c^2_{k_i,l_t,t}} + \delta_i \overline{\mathbb{E} \| \nabla \bar{f}_i(x_{i,t-1}) \|^2} + \frac{\rho}{\Delta} D^2 \overline{\mathbb{E} \| \nabla f_i(x_{i,t-1}) \|^2} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} \right).
\end{aligned}
$$

Justifications for the labeled steps are:

- (i),(ii),(iii) Young's inequality
- (iv) First, we can can interchange the sum and the norm due to independent stochasticity of the gradients. Then by the Tower Property and Law of Total Variance for the 1st and 3rd steps respectively,

$$\mathbb{E} \| \mathbb{E}_x g_j(x_{i,t-1}) - \mathbb{E} g_j(x_{i,t-1})) \|^2 = \mathbb{E} \| \mathbb{E}_x g_j(x_{i,t-1}) - \mathbb{E}[\mathbb{E}_x g_j(x_{i,t-1})] \|^2$$

$$
\begin{aligned}
&= \mathrm{Var}(\mathbb{E}_x(g_j(x_{i,t-1}))) \\
&= \mathrm{Var}(g_j(x_{i,t-1})) - \mathbb{E}(\mathrm{Var}_x(g_j(x_{i,t-1}))) \\
&\leq \mathrm{Var}(g_j(x_{i,t-1})) - \mathbb{E}\|g_j(x_{i,t-1}) - \mathbb{E}_x g_j(x_{i,t-1})\|^2 \\
&\lesssim \rho^2,
\end{aligned}
$$

where the last inequality follows since the two terms above it are both bounded by ρ^2.

- (v) We prove this bound in Lemmas 10, 11, and 12. It shows that, from point i's perspective, the error of its cluster-center-estimate is composed of three quantities: $\mathcal{T}_1$, the error introduced by our thresholding procedure on the good points which belong to i's cluster (and therefore ideally are included within the thresholding radius); $\mathcal{T}_2$, which accounts for the variance of the clipped points in i's cluster; and $\mathcal{T}_3$, the error due to the good points which don't belong to i's cluster (and therefore ideally are forced outside the thresholding radius).
- (vi) $(1 + {}^{x}\!/_{2})^2(1-x)^2 \leq 1 - x$ for all $x \in [0, 1]$

□

Lemma 10 (Bound $\mathcal{T}_1$: Error due to In-Cluster Good Points)**.**

$$
\overline{\mathbb{E}_{j \in \mathcal{G}_i : j \sim i} \|\mathbb{E}(y_{j,l_t,t} - g_j(x_{i,t-1}))\|^2} \lesssim \overline{c^2_{k_i,l_t,t}} + \overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho^3}{\delta_i \Delta}.
$$

Proof of Lemma 10. In this sequence of steps, we bound the clustering error due to good points from client i's cluster. By definition of $y_{j,l_t,t}$,

$$
\begin{aligned}
\mathbb{E}\|y_{j,l_t,t} - g_j(x_{i,t-1})\| &= \mathbb{E}[\|v_{i,l_t-1,t} - g_j(x_{i,t-1})\| \mathbb{1}(\|v_{i,l_t-1,t} - g_j(x_{i,t-1})\| > \tau_{k_i,l_t,t})] \\
&\leq \frac{\mathbb{E}[\|v_{i,l_t-1,t} - g_j(x_{i,t-1})\|^2 \mathbb{1}(\|v_{i,l_t-1,t} - g_j(x_{i,t-1})\| > \tau_{k_i,l_t,t})]}{\tau_{k_i^t,l_t}} \\
&\leq \frac{\mathbb{E}\|v_{i,l_t-1,t} - g_j(x_{i,t-1})\|^2}{\tau_{k_i,l_t,t}}.
\end{aligned}
$$

Therefore, by Jensen's inequality and plugging in the value for $\tau_{k_i,l_t,t}$,

$$
\begin{aligned}
&\|\mathbb{E}(y_{j,l_t,t} - g_j(x_{i,t-1}))\|^2 \\
&\leq (\mathbb{E}\|y_{j,l_t,t} - g_j(x_{i,t-1})\|)^2 \\
&\leq \frac{(\mathbb{E}\|v_{i,l_t-1,t} - g_j(x_{i,t-1})\|^2)^2}{\tau^2_{k_i,l_t,t}} \\
&\lesssim \left(\frac{1}{\tau^2_{k_i,l_t,t}}\right) \cdot \Bigg(\mathbb{E}\|v_{i,l_t-1,t} - \mathbb{E}_x \bar{g}_i(x_{i,t-1})\|^2 + \mathbb{E}\|\mathbb{E}_x \bar{g}_i(x_{i,t-1}) - \mathbb{E}_x g_j(x_{i,t-1})\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \mathbb{E}\|g_j(x_{i,t-1}) - \mathbb{E}_x g_j(x_{i,t-1})\|^2)^2\Big)^2 \\
&\le \frac{(c_{k_i,l_t,t}^2 + A^2\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2 + \rho^2)^2}{\tau_{k_i,l_t,t}^2} \\
&\lesssim \frac{(c_{k_i,l_t,t}^2 + A^2\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2 + \rho^2)^2}{c_{k_i,l_t,t}^2 + A^4\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2 + \delta_i\rho\Delta} \\
&\lesssim \left(1 + \frac{\rho}{\delta_i\Delta}\right)c_{k_i,l_t,t}^2 + \left(1 + \frac{\rho A^2}{\delta_i\Delta}\right)\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2 + \frac{\rho^3}{\delta_i\Delta} \\
&\lesssim c_{k_i,l_t,t}^2 + \mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2 + \frac{\rho^3}{\delta_i\Delta},
\end{aligned}
\tag{2.24}
$$

where the last inequality follows from constraints on Δ. The second inequality follows from Young's inequality. The second-to-last inequality follows by separating the fraction into a sum of fractions and selecting terms from the denominator for each fraction that cancel with terms in the numerator to achieve the desired rate.

Summing (2.24) over t and dividing by T, we have

$$
\begin{aligned}
\overline{\mathbb{E}_{j\in\mathcal{G}_i:j\sim i}\|\mathbb{E}_x(y_{j,l_t,t} - g_j(x_{i,t-1}))\|^2} &= \overline{\mathbb{E}\|y_{j,l_t,t} - g_j(x_{i,t-1})\|^2} \\
&\lesssim \overline{c_{k_i,l_t,t}^2} + \overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} + \frac{\rho^3}{\delta_i\Delta}.
\end{aligned}
$$

□

Lemma 11 (Bound $\mathcal{T}_2$: Variance of Clipped Points)**.**

$$
\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t})\right\|^2 \le \frac{n_i}{|\mathcal{G}_i|^2}\rho^2.
$$

Proof of Lemma 11. The first thing to note is that the elements in the sum $\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t})$ are not independent. Therefore, we cannot get rid of the cross terms when expanding the squared-norm. However, if for each round of thresholding we sampled a fresh batch of points to set the new cluster-center estimate, then the terms would be independent. With such a resampling strategy, our bounds in these proofs only change by a constant factor. Therefore, for ease of analysis, we will assume the terms in the sum are independent. In that case,

$$
\begin{aligned}
\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t})\right\|^2 &\le \frac{n_i}{|\mathcal{G}_i|^2}\mathbb{E}\|y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t}\|^2 \\
&\le \frac{n_i}{|\mathcal{G}_i|^2}\mathbb{E}\|g_j(x_{i,t-1}) - \mathbb{E}g_j(x_{i,t-1})\|^2
\end{aligned}
$$

$$\leq \frac{n_i}{|\mathcal{G}_i|^2}\rho^2,$$

where the second-to-last inequality follows from the contractivity of the thresholding procedure. □

Lemma 12 (Bound $\mathcal{T}_3$: Error due to Out-of-Cluster Good Points)**.**

$$\begin{aligned}\overline{\mathbb{E}_{j\in\mathcal{G}_i: j\not\sim i}\|y_{j,l_t,t} - \mathbb{E}_x \bar{g}_i(x_{i,t-1})\|^2} \lesssim \left(1+\frac{\delta_i}{2}\right)\overline{c^2_{k_i,l_t,t}} + \delta_i\overline{\mathbb{E}\|\nabla \bar{f}_i(x_{i,t-1})\|^2} \\ + \frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2} + \frac{\rho^3}{\Delta}.\end{aligned}$$

Proof of Lemma 12. This sequence of steps bounds the clustering error due to points not from client i's cluster. Using Young's inequality for the first step,

$$\begin{aligned}
&\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2\\
&\leq\left(1+\frac{\delta_i}{2}\right)\mathbb{E}\|v_{i,l_t-1,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2+\left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-v_{i,l_t-1,t}\|^2\\
&\leq\left(1+\frac{\delta_i}{2}\right)c^2_{k_i,l_t,t}+\left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-v_{i,l_t-1,t}\|^2\\
&=\left(1+\frac{\delta_i}{2}\right)c^2_{k_i,l_t,t}\\
&\qquad+\left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}[\|g_j(x_{i,t-1})-v_{i,l_t-1,t}\|^2\mathbb{1}\{\|g_j(x_{i,t-1})-v_{i,l_t-1,t}\|\leq\tau_{k_i,l_t,t}\}]\\
&\leq\left(1+\frac{\delta_i}{2}\right)c^2_{k_i,l_t,t}+\left(1+\frac{2}{\delta_i}\right)\tau^2_{k_i,l_t,t}\mathbb{P}_{j\in\mathcal{G}_i:j\not\sim i}(\|g_j(x_{i,t-1})-v_{i,l_t-1,t}\|\leq\tau_{k_i,l_t,t})\\
&\lesssim\left(1+\frac{\delta_i}{2}\right)c^2_{k_i,l_t,t}\\
&\qquad+\left(\frac{1}{\delta_i}\right)(c^2_{k_i,l_t,t}+A^4\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2+\delta_i\rho\Delta)\mathbb{P}_{j\in\mathcal{G}_i:j\not\sim i}(\|g_j(x_{i,t-1})-v_{i,l_t-1,t}\|\leq\tau_{k_i,l_t,t}).
\end{aligned}$$

The next step is to bound the probability term in the inequality above. Note that if $\|v_{i,l_t-1,t}-g_j(x_{i,t-1})\|\leq\tau_{k_i^t,l_t}$, then

$$\begin{aligned}
\|\mathbb{E}_xg_j(x_{i,t-1})-\mathbb{E}_xg_i(x_{i,t-1})\|^2&\lesssim\|g_j(x_{i,t-1})-\mathbb{E}_xg_j(x_{i,t-1})\|^2+\|g_j(x_{i,t-1})-v_{i,l_t-1,t}\|^2\\
&\qquad+\|v_{i,l_t-1,t}-\mathbb{E}_xg_i(x_{i,t-1})\|^2\\
&\lesssim\|g_j(x_{i,t-1})-\mathbb{E}_xg_j(x_{i,t-1})\|^2+\tau^2_{k_i,l_t,t}\\
&\qquad+\|v_{i,l_t-1,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2+\|\mathbb{E}_xg_i(x_{i,t-1})-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2\\
&\lesssim\|g_j(x_{i,t-1})-\mathbb{E}_xg_j(x_{i,t-1})\|^2+\tau^2_{k_i,l_t,t}\\
&\qquad+\|v_{i,l_t-1,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2+A^2\|\nabla\bar{f}_i(x_{i,t-1})\|^2.
\end{aligned}$$

By Assumption 4, the previous inequality implies

$$\begin{aligned}
\Delta^2-D^2\|\nabla f_i(x_{i,t-1})\|^2\lesssim\|g_j(x_{i,t-1})-\mathbb{E}_xg_j(x_{i,t-1})\|^2+\tau^2_{k_i,l_t,t}+A^2\|\nabla\bar{f}_i(x_{i,t-1})\|^2\\
+\|v_{i,l_t-1,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2
\end{aligned}$$

which, summing over t and dividing by T, implies

$$\begin{aligned}
\Delta^2-\overline{\tau^2_{k_i,l_t,t}}\lesssim\overline{\|g_j(x_{i,t-1})-\mathbb{E}_xg_j(x_{i,t-1})\|^2}+\overline{\|v_{i,l_t-1,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2}\\
+A^2\overline{\|\nabla\bar{f}_i(x_{i,t-1})\|^2}+D^2\overline{\|\nabla f_i(x_{i,t-1})\|^2}.
\end{aligned}$$

By Markov's inequality, the probability of this event is upper-bounded by

$$\frac{\rho^2+\overline{\mathbb{E}\|v_{i,l_t-1,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2}+A^2\overline{\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2}+D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2}}{\Delta^2-\overline{\tau^2_{k_i,l_t,t}}}$$
$$\lesssim\frac{\rho^2+\overline{c^2_{k_i,l_t,t}}+A^2\overline{\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2}+D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2}}{\Delta^2},$$

where the second inequality holds due to the constraint on Δ from the theorem statement. Therefore,

$$\begin{aligned}
&\overline{\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2}\\
&\lesssim\left(1+\frac{\delta_i}{2}+\frac{\rho}{\Delta}+\frac{\rho^2+\overline{c^2_{k_i,l_t,t}}+A^2\overline{\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2}+D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2}}{\delta_i\Delta^2}\right)\overline{c^2_{k_i,l_t,t}}\\
&\quad+\left(\frac{\rho A^2}{\Delta}+\frac{A^4(\rho^2+\overline{c^2_{k_i,l_t,t}}+A^2\overline{\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2}+D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2})}{\delta_i\Delta^2}\right)\overline{\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2}\\
&\quad+\frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2}+\frac{\rho^3}{\Delta}\\
&\lesssim\left(1+\frac{\delta_i}{2}\right)\overline{c^2_{k_i,l_t,t}}+\delta_i\overline{\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2}+\frac{\rho}{\Delta}D^2\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2}+\frac{\rho^3}{\Delta},
\end{aligned}$$

where for the last inequality we again apply the constraint on Δ. □

Lemma 13 (Clustering Error due to Bad Points)**.**

$$\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i}y_{j,l_t,t}\right)-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\right\|^2}\lesssim\overline{c^2_{k_i,l_t,t}}+A^4\overline{\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2}+\delta_i\rho\Delta$$

Proof of Lemma 13. This lemma bounds the clustering error due to the bad clients from client i's perspective. The goal of such clients would be to corrupt the cluster-center estimate of client i's cluster as much as possible at each round. They can have the maximum negative effect by setting their gradients to be just inside the thresholding radius around client i's cluster-center estimate. This way, the gradients will keep their value (rather than be assigned the value of the current cluster-center estimate per our update rule), but they will have maximal effect in moving the cluster-center estimate from its current position. Therefore, in step 3 of the inequalities below, we can not do better than bounding the distance between these bad points and the current cluster center estimate (i.e. $\|y_{j,l_t,t}-v_{i,l_t-1,t}\|^2$) by the thresholding radius ($\tau^2_{k_i,l_t,t}$).

$$\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i}y_{j,l_t,t}\right)-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\right\|^2\leq\mathbb{E}_{j\in\mathcal{B}_i}\|y_{j,l_t,t}-\mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2$$

$$
\begin{aligned}
&\lesssim \mathbb{E}_{j\in\mathcal{B}_i}\|y_{j,l_t,t} - v_{i,l_t-1,t}\|^2 + \mathbb{E}\|v_{i,l_t-1,t} - \mathbb{E}_x\bar{g}_i(x_{i,t-1})\|^2 \\
&\le \tau^2_{k_i,l_t,t} + c^2_{k_i,l_t,t} \\
&\lesssim c^2_{k_i,l_t,t} + A^4\mathbb{E}\|\nabla\bar{f}_i(x_{i,t-1})\|^2 + \delta_i\rho\Delta.
\end{aligned}
$$

The last inequality applies the definition of $\tau_{k_i,l_t,t}$, and the result of the lemma follows by summing this inequality over t and dividing by T. □

Proof of Theorem 4

First we establish some notation.

Notation.

- $\mathcal{G}_i$ are the good clients and $\mathcal{B}_i$ the bad clients from client i's perspective.
- $\mathbb{E}_x$ denotes conditional expectation given the parameter, e.g. $\mathbb{E}_x g(x) = \mathbb{E}[g(x)|x]$. $\mathbb{E}$ denotes expectation over all randomness.
- k_i^t is the cluster to which client i is assigned at round t of the algorithm.
- $\overline{X_t} \triangleq \frac{1}{T}\sum_{t=1}^T X_t$ for a general variable X_t indexed by t.
- $\bar{f}_i(x) \triangleq \frac{1}{n_i}\sum_{j\in\mathcal{G}_i:j\sim i} f_j(x)$
- $\bar{m}_{i,t} \triangleq \frac{1}{n_i}\sum_{j\in\mathcal{G}_i:j\sim i} m_{j,t}$
- We introduce a variable ρ^2 to bound the variance of the momentums

  $$\mathbb{E}_x\|m_{i,t} - \mathbb{E}_x m_{i,t}\|^2 \le \rho^2,$$

  and show in Lemma 14 how this can be written in terms of the variance of the gradients, σ^2.
- l_t is the number of rounds that Threshold-Clustering is run in round t of Federated-Clustering.
- k_i denotes the cluster to which client i is assigned.
- $v_{k_i,l,t}$ denotes the gradient update for client i in round t of Momentum-Clustering and round l of Threshold-Clustering. That is, $v_{k_i,l,t}$ corresponds to the quantity returned in Step 4 of Algorithm 4.

- To facilitate the proof, we introduce a variable $c_{k_i,l,t}$ that quantifies the distance from the cluster-center-estimates to the true cluster means. Specifically, for client i's cluster k_i at round t of Federated-Clustering and round l of Threshold-Clustering we set

$$c^2_{k_i,l,t} = \mathbb{E}\|v_{k_i,l-1,t} - \mathbb{E}_x \bar{m}_{i,t}\|^2.$$

- For client i's cluster k_i at round t of Federated-Clustering and round l of Threshold-Clustering, we use thresholding radius

$$\tau^2_{k_i,l,t} \approx c^2_{k_i,l,t} + \delta_i \rho \Delta.$$

- Finally, we introduce a variable, $y_{j,l,t}$, to denote the points clipped by Threshold-Clustering:

$$v_{k_i,l,t} = \frac{1}{N} \sum_{j \in [N]} \underbrace{\mathbb{1}(\|m_{j,t} - v_{k_i,l-1,t}\| \leq \tau_{k_i,l,t}) + v_{k_i,l-1,t}\mathbb{1}(\|m_{j,t} - v_{k_i,l-1,t}\| > \tau_{k_i,l,t})}_{y_{j,l,t}}.$$

Proof of Theorem 4. In this proof, our goal is to bound $\overline{\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2}$. We use L-smoothness of the loss objectives to get started, and justify the non-trivial steps afterwards.

$$\begin{aligned}
\mathbb{E} f_i(x_{i,t}) &\overset{(i)}{\leq} \mathbb{E} f_i(x_{i,t-1}) + \mathbb{E}\langle \nabla f_i(x_{i,t-1}), x_{i,t} - x_{i,t-1}\rangle + \frac{L}{2}\mathbb{E}\|x_{i,t} - x_{i,t-1}\|^2 \\
&= \mathbb{E} f_i(x_{i,t-1}) - \eta \mathbb{E}\langle \nabla f_i(x_{i,t-1}), v_{k_i,l_t,t}\rangle + \frac{L\eta^2}{2}\mathbb{E}\|v_{k_i,l_t,t}\|^2 \\
&= \mathbb{E} f_i(x_{i,t-1}) + \frac{\eta}{2}\mathbb{E}\|v_{k_i,l_t,t} - \nabla f_i(x_{i,t-1})\|^2 \\
&\qquad - \frac{\eta}{2}\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2 - \frac{\eta}{2}(1 - L\eta)\mathbb{E}\|v_{k_i,l_t,t}\|^2 \\
&\overset{(ii)}{\lesssim} \mathbb{E} f_i(x_{i,t-1}) + \eta \mathbb{E}\|v_{k_i,l_t,t} - \mathbb{E}_x \bar{m}_{i,t}\|^2 + \eta \mathbb{E}\|\mathbb{E}_x \bar{m}_{i,t} - \nabla f_i(x_{i,t-1})\|^2 \\
&\qquad - \frac{\eta}{2}\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2 - \frac{\eta}{2}(1 - L\eta)\mathbb{E}\|v_{k_i,l_t,t}\|^2 \\
&\overset{(iii)}{\lesssim} \mathbb{E} f_i(x_{i,t-1}) + \eta\left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i \rho \Delta\right) + \eta \mathbb{E}\|\mathbb{E}_x \bar{m}_{i,t} - \nabla f_i(x_{i,t-1})\|^2 \\
&\qquad - \frac{\eta}{2}\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2 - \frac{\eta}{2}(1 - L\eta)\mathbb{E}\|v_{k_i,l_t,t}\|^2.
\end{aligned} \tag{2.25}$$

Justifications for the labeled steps are:

- (i) L-smoothness of f_i and $\eta \leq {}^1\!/L$
- (ii) Young's inequality
- (iii) Lemma 15

Now it remains to bound the $\mathbb{E}\|\mathbb{E}_x\bar{m}_{i,t} - \nabla f_i(x_{i,t-1})\|^2$ term.

$$\begin{aligned}
\mathbb{E}\|\mathbb{E}_x\bar{m}_{i,t} - \nabla f_i(x_{i,t-1})\|^2 &\leq \mathbb{E}\|\mathbb{E}_x m_{i,t} - \nabla f_i(x_{i,t-1})\|^2 \\
&= (1-\alpha)^2 \cdot \\
&\quad \mathbb{E}\|\mathbb{E}_x m_{i,t-1} - \nabla f_i(x_{i,t-2}) + \nabla f_i(x_{i,t-2}) - \nabla f_i(x_{i,t-1})\|^2 \\
&\overset{(i)}{\lesssim} (1-\alpha)^2(1+\alpha)\mathbb{E}\|\mathbb{E}_x m_{i,t-1} - \nabla f_i(x_{i,t-2})\|^2 \\
&\quad + (1-\alpha)^2\left(1+\frac{1}{\alpha}\right)L^2\eta^2\mathbb{E}\|v_{k_i,l_{t-1},t-1}\|^2 \\
&\overset{(ii)}{\leq} \frac{1}{2}(1-L\eta)\sum_{t=2}^{T}\mathbb{E}\|v_{k_i,l_{t-1},t-1}\|^2,
\end{aligned}$$

where justifications for the labeled steps are:

- (i) Young's inequality
- (ii) Assumption that $\alpha \gtrsim L\eta$

Plugging this bound back into (2.25), summing over $t = 1 : T$, and dividing by T gives

$$\frac{\eta}{2T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2 \lesssim \frac{\mathbb{E}(f_i(x_{i,0}) - f_i^*)}{T} + \eta\left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right). \tag{2.26}$$

By the variance reduction from momentum (Lemma 14) it follows from (2.26) that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2 &\lesssim \frac{\mathbb{E}(f_i(x_{i,0}) - f_i^*)}{\eta T} + \left(\frac{\alpha\sigma^2}{n_i} + \frac{\alpha^{3/2}\sigma^3}{\Delta} + \beta_i\sqrt{\alpha}\sigma\Delta\right) \\
&\lesssim \frac{\mathbb{E}(f_i(x_{i,0}) - f_i^*)}{\eta T} + \left(\frac{L\eta\sigma^2}{n_i} + \frac{L\eta\sigma^3}{\Delta} + \beta_i\sqrt{L\eta}\sigma\Delta\right).
\end{aligned}$$

Finally, setting $\eta \lesssim \min\left\{\frac{1}{L}, \sqrt{\frac{\mathbb{E}(f_i(x_{i,0})-f_i^*)}{LT(\sigma^2/n_i+\sigma^3/\Delta)}}\right\}$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_i(x_{i,t-1})\|^2 \lesssim \sqrt{\frac{\sigma^2/n_i + \sigma^3/\Delta}{T}} + \frac{\beta_i\sigma\Delta}{T^{\frac{1}{4}}(\sigma^2/n_i + \sigma^3/\Delta)^{\frac{1}{4}}}.$$

□

Lemma 14 (Variance reduction using Momentum)**.** *Suppose that for all $i \in [N]$ and x,*

$$\mathbb{E}\|g_i(x) - \mathbb{E}_x g_i(x)\|^2 \leq \sigma^2.$$

Then

$$\mathbb{E}\|m_{i,t} - \mathbb{E}_x m_{i,t}\|^2 \leq \alpha\sigma^2.$$

Proof of Lemma 14.

$$\begin{aligned}
\mathbb{E}\|m_{i,t} - \mathbb{E}m_{i,t}\|^2 &= \mathbb{E}\|\alpha(g_i(x_{i,t-1}) - \nabla f_i(x_{i,t-1})) + (1-\alpha)(m_{i,t-1} - \mathbb{E}m_{i,t-1})\|^2 \\
&\leq \alpha^2\mathbb{E}\|g_i(x_{i,t-1}) - \nabla f_i(x_{i,t-1})\|^2 + (1-\alpha)^2\mathbb{E}\|m_{i,t-1} - \mathbb{E}m_{i,t-1}\|^2 \\
&\leq \alpha^2\mathbb{E}\|g_i(x_{i,t-1}) - \nabla f_i(x_{i,t-1})\|^2 + (1-\alpha)\mathbb{E}\|m_{i,t-1} - \mathbb{E}m_{i,t-1}\|^2 \\
&\leq \alpha^2\sigma^2 \sum_{q=0}^{t-1}(1-\alpha)^q \\
&\leq \alpha^2\sigma^2 \frac{(1-\alpha)^t - 1}{(1-\alpha) - 1} \\
&\leq \alpha^2\sigma^2 \frac{1}{\alpha} \\
&= \alpha\sigma^2.
\end{aligned}$$

□

Lemma 15 (Bound on Clustering Error)**.**

$$\overline{\mathbb{E}\|v_{k_i^t,l_t} - \mathbb{E}_x\bar{m}_{i,t}\|^2} \lesssim \frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta$$

Proof of Lemma 15. We prove the main result, and then justify each step afterwards.

$$\begin{aligned}
\overline{\mathbb{E}\|v_{k_i,l_t,t} - \mathbb{E}_x\bar{m}_{i,t}\|^2} &= \overline{\mathbb{E}\left\|\frac{1}{N}\sum_{j\in[N]} y_{j,l_t,t} - \mathbb{E}_x\bar{m}_{i,t}\right\|^2} \\
&= \overline{\mathbb{E}\left\|(1-\beta_i)\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t} - \mathbb{E}_x\bar{m}_{i,t}\right) + \beta_i\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,t,l} - \mathbb{E}_x\bar{m}_{i,t}\right)\right\|^2} \\
&\overset{(i)}{\leq} (1+\beta_i)(1-\beta_i)^2\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t}\right) - \mathbb{E}_x\bar{m}_{i,t}\right\|^2} \\
&\quad + \left(1+\frac{1}{\beta_i}\right)\beta_i^2\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l_t,t}\right) - \mathbb{E}_x\bar{m}_{i,t}\right\|^2} \\
&\lesssim \overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t}\right) - \mathbb{E}_x\bar{m}_{i,t}\right\|^2} + \beta_i\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l_t,t}\right) - \mathbb{E}_x\bar{m}_{i,t}\right\|^2} \\
&\overset{(ii)}{\lesssim} (1-\delta_i+\beta_i)\overline{c^2_{k_i,l_t,t}} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right) \\
&\overset{(iii)}{\lesssim} (1-{\delta_i}/{2})\overline{c^2_{k_i,l_t,t}} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(iv)}{\lesssim} \overline{(1-{\delta_i}/{2})^{l_t} c^2_{k_i,1,t}} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right) \\
&\overset{(v)}{\leq} \frac{\rho}{\Delta}\overline{c^2_{k_i,1,t}} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta\right).
\end{aligned} \tag{2.27}
$$

We justify each step.

- (i) Young's inequality
- (ii) We prove this bound in Lemmas 16 and 20. Importantly, it shows that the clustering error is composed of two quantities: $\mathcal{E}_1$, the error contributed by good points from the cluster's perspective, and $\mathcal{E}_2$, the error contributed by the bad points from the cluster's perspective.
- (iii) Assumption that $\beta_i \lesssim \min(\delta_i, {\delta_i}/{A^4})$
- (iv) Since $\mathbb{E}\|v_{i,l_t,t} - \mathbb{E}_x\bar{m}_{i,t}\|^2 = c^2_{k_i,l_t+1,t}$, the inequality forms a recursion which we unroll over l_t steps.
- (v) Assumption that $l_t \geq \max(1, \frac{\log({\rho}/{\Delta})}{\log(1-{\delta_i}/{2})})$

Finally, we note that

$$
\begin{aligned}
c^2_{k_i^t,1,t} &= \mathbb{E}\|\bar{m}_{i,t} - \mathbb{E}_x m_{i,t}\|^2 \\
&\lesssim \mathbb{E}\|m_{j,t} - \mathbb{E}_x m_{j,t}\|^2 + \mathbb{E}\|\mathbb{E}_x m_{j,t} - \mathbb{E}_x m_{i,t}\|^2 \\
&\leq \rho^2.
\end{aligned}
$$

Applying this bound to (2.27) gives

$$
\overline{\mathbb{E}\|v_{k_i,l_t,t} - \mathbb{E}_x\bar{m}_{i,t}\|^2} \lesssim \frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} + \beta_i\rho\Delta
$$

□

Lemma 16 (Clustering Error due to Good Points)**.**

$$
\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t}\right) - \mathbb{E}_x\bar{m}_{i,t}\right\|^2} \lesssim (1-\delta_i)\overline{c^2_{k_i,l_t,t}} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta}\right).
$$

Proof of Lemma 16. We state the main sequence of steps and then justify them afterwards.

$$
\mathbb{E}\left\|\left(\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i} y_{j,l_t,t}\right) - \mathbb{E}_x\bar{m}_{i,t})\right\|^2
$$

$$
\begin{aligned}
&= \mathbb{E}\left\| \left(\frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i} y_{j,l_t,t} \right) - \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} \mathbb{E}_x m_{j,t} - \left(\frac{1}{n_i} - \frac{1}{|\mathcal{G}_i|} \right) \sum_{j \in \mathcal{G}_i : j \sim i} \mathbb{E}_x m_{j,t} \right\|^2 \\
&= \mathbb{E}\left\| \left(\frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (y_{j,l_t,t} - \mathbb{E}_x m_{j,t}) \right) + \left(\frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \not\sim i} (y_{j,l_t,t} - \mathbb{E}_x \bar{m}_{i,t}) \right) \right\|^2 \\
&\overset{(i)}{\leq} \left(1 + \frac{2}{\delta_i}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (y_{j,l_t,t} - \mathbb{E}_x m_{j,t}) \right\|^2 + \left(1 + \frac{\delta_i}{2}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \not\sim i} (y_{j,l_t,t} - \mathbb{E}_x \bar{m}_{i,t}) \right\|^2 \\
&\overset{(ii)}{\lesssim} \left(1 + \frac{2}{\delta_i}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} \mathbb{E} y_{j,l_t,t} - \mathbb{E}_x m_{j,t} \right\|^2 + \left(1 + \frac{2}{\delta_i}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (y_{j,l_t,t} - \mathbb{E}_x y_{j,l_t,t}) \right\|^2 \\
&\qquad + \left(1 + \frac{\delta_i}{2}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \not\sim i} (y_{j,l_t,t} - \mathbb{E}_x \bar{m}_{i,t}) \right\|^2 \\
&\overset{(iii)}{\lesssim} \left(1 + \frac{2}{\delta_i}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} \mathbb{E}(y_{j,l_t,t} - m_{j,t}) \right\|^2 + \left(1 + \frac{2}{\delta_i}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (\mathbb{E} m_{j,t} - \mathbb{E}_x m_{j,t}) \right\|^2 \\
&\qquad + \left(1 + \frac{2}{\delta_i}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (y_{j,l_t,t} - \mathbb{E} y_{j,l_t,t}) \right\|^2 + \left(1 + \frac{\delta_i}{2}\right) \mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \not\sim i} (y_{j,l_t,t} - \mathbb{E}_x \bar{m}_{i,t}) \right\|^2 \\
&\overset{(iv)}{\lesssim} \left(1 + \frac{2}{\delta_i}\right) \delta_i^2 \underbrace{\mathbb{E}_{j \in \mathcal{G}_i : j \sim i} \| \mathbb{E}_x (y_{j,l_t,t} - m_{j,t}) \|^2}_{\mathcal{T}_1} + \left(1 + \frac{2}{\delta_i}\right) \frac{n_i}{|\mathcal{G}_i|^2} \rho^2 \\
&\qquad + \left(1 + \frac{2}{\delta_i}\right) \frac{n_i}{|\mathcal{G}_i|^2} \underbrace{\mathbb{E}\left\| \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i : j \sim i} (y_{j,l_t,t} - \mathbb{E} y_{j,l_t,t}) \right\|^2}_{\mathcal{T}_2} \\
&\qquad + \left(1 + \frac{\delta_i}{2}\right) (1 - \delta_i)^2 \underbrace{\mathbb{E}_{j \in \mathcal{G}_i : j \not\sim i} \| y_{j,l_t,t} - \mathbb{E}_x \bar{m}_{i,t} \|^2}_{\mathcal{T}_3} \\
&\overset{(v)}{\lesssim} \delta_i \mathbb{E}_{j \in \mathcal{G}_i : j \sim i} \| \mathbb{E}_x (y_{j,l_t,t} - m_{j,t}) \|^2 + \frac{\rho^2}{n_i} + \left(1 + \frac{\delta_i}{2}\right) (1 - \delta_i)^2 \mathbb{E}_{j \in \mathcal{G}_i : j \not\sim i} \| y_{j,l_t,t} - \mathbb{E}_x \bar{m}_{i,t} \|^2 \\
&\overset{(vi)}{\lesssim} \delta_i \left(\overline{c^2_{k_i,l_t,t}} + \frac{\rho^3}{\delta_i \Delta} \right) + \frac{\rho^2}{n_i} + \left(1 + \frac{\delta_i}{2}\right) (1 - \delta_i)^2 \left(\left(1 + \frac{\delta_i}{2}\right) \overline{c^2_{k_i,l_t,t}} + \frac{\rho^3}{\Delta} \right) \\
&\lesssim \delta_i \left(\overline{c^2_{k_i,l_t,t}} + \frac{\rho^3}{\delta_i \Delta} \right) + \frac{\rho^2}{n_i} + (1 - \delta_i) \left(\overline{c^2_{k_i,l_t,t}} + \frac{\rho^3}{\Delta} \right) \\
&\lesssim (1 - \delta_i) \overline{c^2_{k_i,l_t,t}} + \left(\frac{\rho^2}{n_i} + \frac{\rho^3}{\Delta} \right).
\end{aligned}
$$

Justifications for the labeled steps are:

- (i), (ii), (iii) Young's inequality

- (iv) First, we can can interchange the sum and the norm due to independent stochasticity of the momentums. Then, by the Tower Property and Law of Total Variance for the 1st and 3rd steps respectively

$$\begin{aligned}\mathbb{E}\|\mathbb{E}_x m_{j,t} - \mathbb{E}m_{j,t}\|^2 &= \mathbb{E}\|\mathbb{E}_x m_{j,t} - \mathbb{E}[\mathbb{E}_x m_{j,t}]\|^2 \\ &= \mathrm{Var}(\mathbb{E}_x m_{j,t}) \\ &= \mathrm{Var}(m_{j,t}) - \mathbb{E}(\mathrm{Var}_x(m_{j,t})) \\ &= \mathrm{Var}(m_{j,t}) - \mathbb{E}\|m_{j,t} - \mathbb{E}_x m_{j,t}\|^2 \\ &\lesssim \rho^2,\end{aligned}$$

  where the last inequality follows since the two terms above it are both bounded by ρ^2.

- (v) We prove this bound in Lemmas 17, 18, and 19. It shows that, from point i's perspective, the error of its cluster-center-estimate is composed of three quantities: $\mathcal{T}_1$, the error introduced by our thresholding procedure on the good points which belong to i's cluster (and therefore ideally are included within the thresholding radius); $\mathcal{T}_2$, which accounts for the variance of the clipped points in i's cluster; and $\mathcal{T}_3$, the error due to the good points which don't belong to i's cluster (and therefore ideally are forced outside the thresholding radius).

- (vi) $(1 + x/2)^2(1-x)^2 \leq 1 - x$ for all $x \in [0,1]$

□

Lemma 17 (Bound $\mathcal{T}_1$: Error due to In-Cluster Good Points)**.**

$$\overline{\mathbb{E}_{j\in\mathcal{G}_i:j\sim i}\|\mathbb{E}(y_{j,l_t,t} - m_{j,t})\|^2} \lesssim \overline{c^2_{k_i,l_t,t}} + \frac{\rho^3}{\delta_i \Delta}.$$

Proof of Lemma 17. By definition of $y_{j,l_t,t}$,

$$\begin{aligned}\mathbb{E}\|y_{j,l_t,t} - m_{j,t}\| &= \mathbb{E}[\|v_{k_i,l_t-1,t} - m_{j,t}\|\mathbb{1}(\|v_{k_i,l_t-1,t} - m_{j,t}\| > \tau_{k_i,l_t,t})] \\ &\leq \frac{\mathbb{E}[\|v_{k_i,l_t-1,t} - m_{j,t}\|^2\mathbb{1}(\|v_{k_i,l_t-1,t} - m_{j,t}\| > \tau_{k_i,l_t,t})]}{\tau_{k_i,l_t,t}}.\end{aligned}$$

Therefore by Jensen's inequality,

$$\begin{aligned}\|\mathbb{E}y_{j,l_t,t} - m_{j,t}\|^2 &\leq (\mathbb{E}\|y_{j,l_t,t} - m_{j,t}\|)^2 \\ &\leq \frac{(\mathbb{E}\|v_{k_i,l_t-1,t} - m_{j,t}\|^2)^2}{\tau^2_{k_i,l_t,t}} \\ &\lesssim \frac{(\mathbb{E}\|v_{k_i,l_t-1,t} - \mathbb{E}_x\bar{m}_{i,t}\|^2 + \mathbb{E}\|\mathbb{E}_x\bar{m}_{i,t} - m_{j,t}\|^2)^2}{\tau^2_{k_i,l_t,t}}\end{aligned}$$

$$
\begin{aligned}
&= \frac{(\mathbb{E}\|v_{k_i,l_t-1,t} - \mathbb{E}_x\bar{m}_{i,t}\|^2 + \mathbb{E}\|\mathbb{E}_x m_{j,t} - m_{j,t}\|^2)^2}{\tau^2_{k_i,l_t,t}} \\
&\le \frac{(c^2_{k_i,l_t,t} + \rho^2)^2}{\tau^2_{k_i,l_t,t}} \\
&\lesssim \frac{(c^2_{k_i,l_t,t} + \rho^2)^2}{c^2_{k_i,l_t,t} + \delta_i\rho\Delta} \\
&\lesssim \left(1 + \frac{\rho}{\delta_i\Delta}\right)c^2_{k_i,l_t,t} + \frac{\rho^3}{\delta_i\Delta} \\
&\lesssim c^2_{k_i,l_t,t} + \frac{\rho^3}{\delta_i\Delta}.
\end{aligned}
$$

Summing this inequality over t and dividing by T, we have

$$
\begin{aligned}
\overline{\mathbb{E}_{j\in\mathcal{G}_i:j\sim i}\|\mathbb{E}_x(y_{j,l_t,t} - m_{j,t})\|^2} &\le \overline{\mathbb{E}\|y_{j,l_t,t} - m_{j,t}\|^2} \\
&\lesssim \overline{c^2_{k_i,l_t,t}} + \frac{\rho^3}{\delta_i\Delta}.
\end{aligned}
$$

□

Lemma 18 (Bound $\mathcal{T}_2$: Variance of Clipped Points)**.**

$$
\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t})\right\|^2 \le \frac{n_i}{|\mathcal{G}_i|^2}\rho^2.
$$

Proof of Lemma 18. Note that the elements in the sum $\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t})$ are not independent. Therefore, we cannot get rid of the cross terms when expanding the squared-norm. However, if for each round of thresholding we sampled a fresh batch of points to set the new cluster-center estimate, then the terms would be independent. With this resampling strategy, our bounds would only change by a constant factor. Therefore, for ease of analysis, we will assume the terms in the sum are independent. In that case,

$$
\begin{aligned}
\mathbb{E}\left\|\frac{1}{|\mathcal{G}_i|}\sum_{j\in\mathcal{G}_i:j\sim i}(y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t})\right\|^2 &\le \frac{n_i}{|\mathcal{G}_i|^2}\mathbb{E}\|y_{j,l_t,t} - \mathbb{E}y_{j,l_t,t}\|^2 \\
&\le \frac{n_i}{|\mathcal{G}_i|^2}\mathbb{E}\|m_{j,t} - \mathbb{E}m_{j,t}\|^2 \\
&\le \frac{n_i}{|\mathcal{G}_i|^2}\rho^2,
\end{aligned}
$$

where the second-to-last inequality follows from the contractivity of the thresholding procedure.

□

Lemma 19 (Bound $\mathcal{T}_3$: Error due to Out-of-Cluster Good Points)**.**

$$\overline{\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-\mathbb{E}_x\bar{m}_{i,t}\|^2}\lesssim\left(1+\frac{\delta_i}{2}\right)\overline{c^2_{k_i,l_t,t}}+\frac{\rho^3}{\Delta}.$$

Proof of Lemma 19. This sequence of steps bounds the clustering error due to points not from client i's cluster. Using Young's inequality for the first step,

$$\begin{aligned}
&\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-\mathbb{E}_x\bar{m}_{i,t}\|^2\\
&\le\left(1+\frac{\delta_i}{2}\right)\mathbb{E}\|v_{k_i^t,l_t-1}-\mathbb{E}_x\bar{m}_{i,t}\|^2+\left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-v_{k_i,l_t-1,t}\|^2\\
&\le\left(1+\frac{\delta_i}{2}\right)c^2_{k_i,l_t,t}+\left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-v_{k_i,l_t-1,t}\|^2\\
&=\left(1+\frac{\delta_i}{2}\right)c^2_{k_i,l_t,t}+\left(1+\frac{2}{\delta_i}\right)\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}[\|m_{j,t}-v_{k_i,l_t-1,t}\|^2\mathbb{1}\{\|m_{j,t}-v_{k_i,l_t-1,t}\|\le\tau_{k_i,l_t,t}\}]\\
&\le\left(1+\frac{\delta_i}{2}\right)c^2_{k_i,l_t,t}+\left(1+\frac{2}{\delta_i}\right)\tau^2_{k_i,l_t,t}\mathbb{P}_{j\in\mathcal{G}_i:j\not\sim i}(\|m_{j,t}-v_{k_i,l_t-1,t}\|\le\tau_{k_i,l_t,t})\\
&\lesssim\left(1+\frac{\delta_i}{2}\right)c^2_{k_i,l_t,t}+\left(\frac{1}{\delta_i}\right)(c^2_{k_i,l_t,t}+\delta_i\rho\Delta)\mathbb{P}_{j\in\mathcal{G}_i:j\not\sim i}(\|m_{j,t}-v_{k_i^t,l_t-1}\|\le\tau_{k_i,l_t,t}).
\end{aligned}$$

If $\|v_{k_i^t,l_t-1}-m_{j,t}\|\le\tau_{k_i,l_t,t}$, then

$$\begin{aligned}
\|\mathbb{E}_x m_{j,t}-\mathbb{E}_x m_{i,t}\|^2&\lesssim\|m_{j,t}-\mathbb{E}_x m_{j,t}\|^2+\|m_{j,t}-v_{k_i,l_t-1,t}\|^2+\|v_{k_i^t,l_t-1}-\mathbb{E}_x\bar{m}_{i,t}\|^2\\
&\lesssim\|m_{j,t}-\mathbb{E}_x m_{j,t}\|^2+\tau^2_{k_i,l_t,t}+\|v_{k_i,l_t-1,t}-\mathbb{E}_x\bar{m}_{i,t}\|^2
\end{aligned}$$

By Assumption 9,

$$\Delta^2\lesssim\|m_{j,t}-\mathbb{E}_x m_{j,t}\|^2+\tau^2_{k_i,l_t,t}+\|v_{k_i,l_t-1,t}-\mathbb{E}_x\bar{m}_{i,t}\|^2$$

which, summing over t and dividing by T, implies

$$\overline{\|m_{j,t}-\mathbb{E}_x m_{j,t}\|^2}+\overline{\|v_{k_i,l_t-1,t}-\mathbb{E}_x\bar{m}_{i,t}\|^2}\gtrsim\Delta^2-\overline{\tau^2_{k_i,l_t,t}}.$$

By Markov's inequality, the probability of this event is upper-bounded by

$$\frac{\rho^2+\overline{\mathbb{E}\|v_{k_i,l_t-1,t}-\mathbb{E}_x\bar{m}_{i,t}\|^2}}{\Delta^2-\overline{\tau^2_{k_i,l_t,t}}}\lesssim\frac{\rho^2+\overline{c^2_{k_i,l_t,t}}}{\Delta^2},$$

where the inequality holds due to the constraint on Δ from the theorem statement. Therefore,

$$\begin{aligned}
\overline{\mathbb{E}_{j\in\mathcal{G}_i:j\not\sim i}\|y_{j,l_t,t}-\mathbb{E}_x\bar{m}_{i,t}\|^2}&\lesssim\left(1+\frac{\delta_i}{2}+\frac{\rho}{\Delta}+\frac{\rho^2+\overline{c^2_{k_i,l_t,t}}}{\delta_i\Delta^2}\right)\overline{c^2_{k_i,l_t,t}}+\frac{\rho^3}{\Delta}\\
&\lesssim\left(1+\frac{\delta_i}{2}\right)\overline{c^2_{k_i,l_t,t}}+\frac{\rho^3}{\Delta},
\end{aligned}$$

where again we apply the constraint on Δ for the second inequality. □

Lemma 20 (Clustering Error due to Bad Points)**.**

$$\overline{\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l_t,t}\right) - \mathbb{E}_x \bar{m}_{i,t}\right\|^2} \lesssim \overline{c^2_{k_i,l_t,t}} + \delta_i \rho \Delta$$

Proof of Lemma 20. This lemma gives a bound on the clustering error due to the bad clients from client i's perspective. The goal of such clients would be to corrupt the cluster-center estimate of client i's cluster as much as possible at each round. They can have the maximum negative effect by setting their gradients to be just inside the thresholding radius around client i's cluster-center estimate. This way, the gradients will keep their value (rather than be assigned the value of the current cluster-center estimate per our update rule), but they will have maximal effect in moving the cluster-center estimate from its current position. Therefore, in step 3 of the inequalities below, we can not do better than bounding the distance between these bad points and the current cluster center estimate (i.e. $\|y_{j,l_t,t} - v_{k_i,l_t-1,t}\|^2$) by the thresholding radius $\tau_{k_i,l_t,t}$.

$$\begin{aligned}
\mathbb{E}\left\|\left(\frac{1}{|\mathcal{B}_i|}\sum_{j\in\mathcal{B}_i} y_{j,l_t,t}\right) - \mathbb{E}_x \bar{m}_{i,t}\right\|^2 &\leq \mathbb{E}_{j\in\mathcal{B}_i}\|y_{j,l_t,t} - \mathbb{E}_x \bar{m}_{i,t}\|^2 \\
&\lesssim \mathbb{E}_{j\in\mathcal{B}_i}\|y_{j,l_t,t} - v_{k_i,l_t-1,t}\|^2 + \mathbb{E}\|v_{k_i,l_t-1,t} - \mathbb{E}_x \bar{m}_{i,t}\|^2 \\
&\leq \tau^2_{k_i,l_t,t} + c^2_{k_i,l_t,t} \\
&\lesssim c^2_{k_i,l_t,t} + \delta_i \rho \Delta.
\end{aligned}$$

The last inequality applies the definition of $\tau_{k_i,l_t,t}$, and the result of the lemma follows by summing this inequality over t and dividing by T. □

Chapter 3

Collaborative Learning among Competitors

3.1 Introduction

When the guarantees of a collaborative learning system are misaligned with the objectives of the learners, it can disincentivize participation and prompt defections. Recent work [34–36] examines the incentives that clients have to participate in or defect from a collaborative learning system. Misalignment between system-wide and client objectives can lead to undesirable outcomes. For example, [37] show that some clients might *free-ride*, burdening other participants in the network with all the training work while contributing nothing themselves. [8–10, 13, 25, 38] show that if there is heterogeneity across clients' data distributions the global model returned by standard collaborative learning protocols might perform poorly for individual clients. To address the misalignment problem, [39] propose an algorithm whose model updates guarantee that client losses degrade sufficiently from step to step to ensure that no client defects (albeit at some cost to the accuracy of the final global model). In this paper, we take an economics-based view of the problem, framing client *utility/revenue* as the determining factor in defection. We frame clients as competitive firms who are selling their models' predictions to consumers and competing for market share. As in the standard collaborative learning protocol, the firms collaboratively train a global model, but if at any point in the process their revenue decreases, they defect from participation.

Motivating Example. Consider two autonomous vehicle companies training self-driving models, each with initial access only to their own training data. Further, suppose their individual training data does not fully reflect the distribution on which the models must perform well at test time. For example, one company might have a lot of urban data and very little rural data and the other company the opposite. Clearly, if these companies combined their models, they could offer safer and better cars to consumers. However, by collaborating they might also lose their competitive advantage in the market, disincentivizing them from participating. Our objective is to design a collaboration scheme such that neither firm loses revenue, thus incentivizing participation.

Our Contributions. We frame the collaborative learning system as a duopoly of competitive firms whose conditions for joining the system are to improve (or at least not lose) revenue, and we show that collaboration is possible under such conditions.

1. We first show surprising outcomes of two possible collaboration schemes. When both firms contribute fully to the collaboration scheme, their model qualities improve maximally but their revenues go to zero. When only the low-quality firm contributes to the collaboration scheme, both firms' model qualities and revenues improve.

2. We next design a defection-free algorithm which allows *both* firms to contribute to the collaborative system without losing revenue at any step.

3. We show that, except in trivial cases, our algorithm converges to the Nash bargaining solution. This is a significant result because we show that even when both firms myopically focus on improving their own revenues, a solution is reached that maximizes the joint surplus of the firms.

Related Work

Collaborative learning allows multiple clients to collaboratively train a global model without transmitting raw data [1]. In this paper, we characterize the participants in a collaborative learning system as market competitors who will defect from collaboration if they lose revenue by participating. Competitive behavior of firms in markets is a well-established field of study in economics (see [40] for an overview). Particularly relevant to our work is competition in oligopolies [41]. As in [42], we structure our problem as a duopoly of competitive firms. One difference is that they incentivize collaboration with revenue sharing between the firms rather than a guarantee of no-revenue-loss as we do in this paper. Also relevant, [43] parameterize the data sharing problem in terms of competition-type (Bertrand [44] or Cournot [41]) between firms, the number of data points each firm has, and the difficulty of the learning task, and give conditions on these parameters under which collaboration is profitable. As we do, they analyze various data sharing schemes, such as full vs partial collaboration, and propose Nash bargaining [45] as a strategy for partial collaboration. However, we additionally propose a federated optimization algorithm for reaching the Nash bargaining solution, guaranteeing no defections.

3.2 Collaborative Learning in an Oligopoly

For the rest of the paper, we frame the collaborative learning system as a duopoly (i.e. two firms), but all results can be extended to an oligopoly of more than two firms.

Our setup is the following. Each firm possesses a model whose qualities are initially differentiated by classification accuracy on a target dataset. That is, one firm's model has low accuracy and the other firm's model has high accuracy on the target dataset. The consumers care about performance on the target distribution, which is different from the firms' individual data distributions. For example, in the autonomous vehicle example above, the target distribution would represent a variety of geographical locations, traffic instances, times of day/night, etc. while the individual distributions would not. Additionally we assume that the firms' individual distributions are complementary, so their combined data is distributed as the target distribution, motivating the benefit of collaboration. Finally, we assume that, prior to collaboration, one firm has better initial model quality than the other (e.g. they have more training resources).

A consumer has one of three options: 1) pay a higher price for the high-quality firm's model, 2) pay a lower price for the low-quality firm's model, or 3) buy neither model. We assume that all consumers would prefer the higher-quality model if the prices of both models were the same – that is, the firms' models are *vertically differentiated.* Consumers would be happiest if both firms collaborated fully since this would give them two maximally good models to choose from, but the initially high-quality firm would have sacrificed revenue in this scenario (we show this formally in Section 3.3), causing it to defect. Based on this, our motivating question is: can we incentivize firms to join the collaboration scheme, thus benefiting consumers, while giving them no reason to defect due to revenue loss at any stage of the training process? We answer this question affirmatively.

In the following section, we formally describe the duopoly model.

Duopoly Model

Notation and Assumptions

1. A consumer's type corresponds to how much they value quality of prediction. We assume that consumer-types are uniformly distributed on $\Theta = [0,1]$, where consumer-type $\theta = 0$ places no value on quality and consumer-type $\theta = 1$ places maximal value on quality.

2. We denote the low-quality firm's loss on its dataset with model parameters $x \in \mathcal{X}$ as $f(x;l) \in [0,1]$ and the high-quality firm's loss on its dataset as $f(x;h) \in [0,1]$. In the collaborative learning process, both firms want to solve the optimization problem

$$x^* = \arg\min_{x\in\mathcal{X}} f(x), \qquad \text{where } f(x) \overset{\text{def}}{=} \frac{f(x;l)+f(x;h)}{2}. \tag{3.1}$$

   That is, each firm wants to find the model which has minimal average loss across both firms' datasets. When the objective (3.1) is evaluated at the firms' models x_l and x_h, we use the shorthand notation

$$f_l \overset{\text{def}}{=} \frac{f(x_l; l) + f(x_l; h)}{2}, \qquad\qquad f_h \overset{\text{def}}{=} \frac{f(x_h; l) + f(x_h; h)}{2}.$$

Finally, we define model qualities $q(x) \overset{\text{def}}{=} 1 - f(x)$, $q_l \overset{\text{def}}{=} 1 - f_l$ and $q_h \overset{\text{def}}{=} 1 - f_h$.

3. Consumers pay prices $p_{l/h} \in [0, \infty)$ for the low/high-quality firm's model $x_{l/h}$, where $p_l \leq p_h$.

Equilibrium Quantities

The following definition gives the consumer's utility.

Definition 2. *[Consumer Utility] A type-θ consumer has utility*

$$U_c(\theta) = \begin{cases} \theta q_h - p_h & \text{if buys high-quality firm's model} \\ \theta q_l - p_l & \text{if buys low-quality firm's model} \\ 0 & \text{if buys neither model.} \end{cases} \tag{3.2}$$

The consumer utilities in Definition 2 induce the following demands for the firms.

Lemma 21 (Consumer Demands)**.** *Given the utilities in Definition 2,*

1. *consumer demand for the low-quality firm is $D_l = \frac{p_h - p_l}{q_h - q_l} - \frac{p_l}{q_l}$, and*
2. *consumer demand for the high-quality firm is $D_h = 1 - \frac{p_h - p_l}{q_h - q_l}$.*

Proof. See Section 3.7. □

Using the consumer demands in Lemma 21, we can define the utilities of the firms.

Definition 3. *[Firm Utility/Revenue] The low/high firm's utility/revenue from selling its model is*

$$U_{l/h}(q_l, q_h, p_l, p_h) = p_{l/h} D_{l/h}. \tag{3.3}$$

At equilibrium, the firms will set prices p_l and p_h that maximize (3.3), yielding price-optimal utilities.

Lemma 22 (Equilibrium Prices and Utilities)**.** *The optimal prices for the low and high firms are*

$$p_l^* = \frac{q_l(q_h - q_l)}{4q_h - q_l}, \qquad p_h^* = \frac{2q_h(q_h - q_l)}{4q_h - q_l},$$

yielding price-optimal utilities

$$U_l(q_l, q_h, p_l^*, p_h^*) = \frac{q_l q_h (q_h - q_l)}{(4q_h - q_l)^2}, \qquad U_h(q_l, q_h, p_l^*, p_h^*) = \frac{4q_h^2(q_h - q_l)}{(4q_h - q_l)^2}. \tag{3.4}$$

Proof. See Section 3.7. □

Going forward, we will use the shorthand $U_{l/h} \overset{\text{def}}{=} U_{l/h}(q_l, q_h, p_l^*, p_h^*)$.

Remark 1. *Since the firms make their pricing decisions simultaneously and compete based on prices, this is the Bertrand model of competition [44]. This is distinct from other forms of oligopolistic competition, such as Cournot competition [41] in which firms compete based on quantity (i.e. the firms independently and simultaneously decide quantities to produce which then determine market price), or Stackelberg competition [46] in which the firms non-independently and sequentially decide quantities to produce.*

The following proposition states how the firms' utilities vary with quality and is key in our analysis going forward.

Proposition 1 (Relationship between utilities and qualities)**.** *For $q_l \leq q_h$,*

1. *U_h is increasing in q_h,*
2. *U_h is decreasing in q_l,*
3. *U_l is increasing in q_h, and*
4. *U_l is increasing in q_l for $q_l \leq \frac{4}{7}q_h$ and decreasing in q_l otherwise.*

Proof. See Section 3.7 □

In the next section, we examine various collaboration schemes between the firms and observe the impact on their revenues and model qualities.

3.3 Collaboration Schemes

To motivate our method, we describe two potential collaboration schemes between competitors that have sub-optimal and non-intuitive outcomes.

Sharing Protocol. As in standard federated learning protocols, we do not assume that the firms transmit their raw data to each other. Instead, firm A shares with firm B by evaluating the loss of firm B's model on firm A's data. Then firm A shares with firm B the loss, or the gradient of the loss, which allows firm B to optimize the objective (3.1). These exchanges can happen either directly between the firms are through a trusted central coordinator.

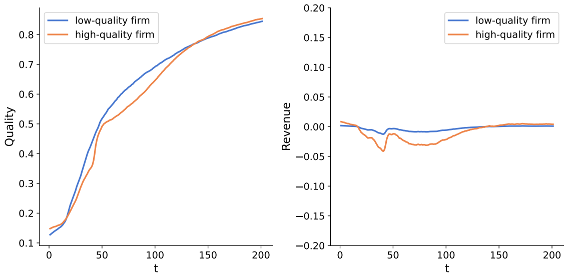


Figure 3.1: Performance of Complete Sharing scheme on MNIST. When both firms share with each other, their models converge to the same utility, driving their utilities to zero.

Notation and Assumptions

1. $f(x; {}^{l}/_{h})$ is convex and L-smooth in x.
2. We use $q_{{}^{l}/_{h},t}$ and $f_{{}^{l}/_{h},t}$ to refer to the firms' objectives when the model parameters are $x_{{}^{l}/_{h},t}$, i.e. the model parameters at round t of optimization.
3. We define $\rho_t = \frac{q_{l,t}}{q_{h,t}}$, the ratio of the firms' model qualities at round t of optimization.
4. We assume model qualities can only improve or stay the same, not degrade.

Complete Collaboration

In this arrangement, both firms fully collaborate, sharing their models with each other and therefore obtaining identical-quality models. (Note that this algorithm is just FedAvg [1].) While this collaboration scheme is optimal for the consumer, giving them the choice of two maximally high-quality models, it drives both firms' utilities to zero. With identical-quality models, each firm will continually undercut the other's price by small amounts to capture the entire market share, eventually reaching equilibrium prices $p_l = p_h = 0$.

Lemma 23 (Firm Revenues under Complete Collaboration)**.** *Under Complete Collaboration, the firms' equilibrium utilities are $U_l = U_h = 0$.*

Figure 3.1 shows that when both firms' qualities increase freely in a Complete Collaboration scheme, their qualities both improve maximally, benefiting the consumer, but their utilities are driven to zero. Therefore, both firms have cause to defect from this collaboration scheme.

One-sided Collaboration

In One-sided Collaboration, one firm shares its model while the other doesn't. There are two possibilities.

Only high-quality firm shares. From Proposition 1, the high-quality firm's revenue increases in q_h but decreases in q_l. Therefore, if the quality of x_h does not increase sufficiently to compensate for the increase in quality of x_l, the high-quality firm will lose revenue, causing it to defect. (In the proof of Proposition 3, we give this increase-threshold precisely.) In our problem setup, the individual firms' training distributions are different than target distribution on which the qualities of their models are evaluated. Therefore, if the low-quality firm benefits from the high-quality firm's model, its performance on the target distribution will outpace the high-quality firm, which is limited to training on its own data. Figure 3.2a gives an example of this outcome. Due to collaboration, the low-quality firm's model out-performs the high-quality firm's model, causing the high-quality firm's revenue to decrease.

Only low-quality firm shares. From Proposition 1, both firms' utilities increase in q_h. Therefore, both firms will increase their revenue if the low-quality firm shares its model with the high-quality firm. Figure 3.2b depicts the outcome of this collaboration scheme. Over time, both firms' revenues increase. While this arrangement is defection-free, the low-quality firm is stuck with its own training data, causing it to potentially have lower revenue that it would under a more equitable scheme. To address this, we next propose a defection-free scheme in which *both* firms participate in collaboration without losing revenue.

3.4 Defection-Free Collaborative Learning

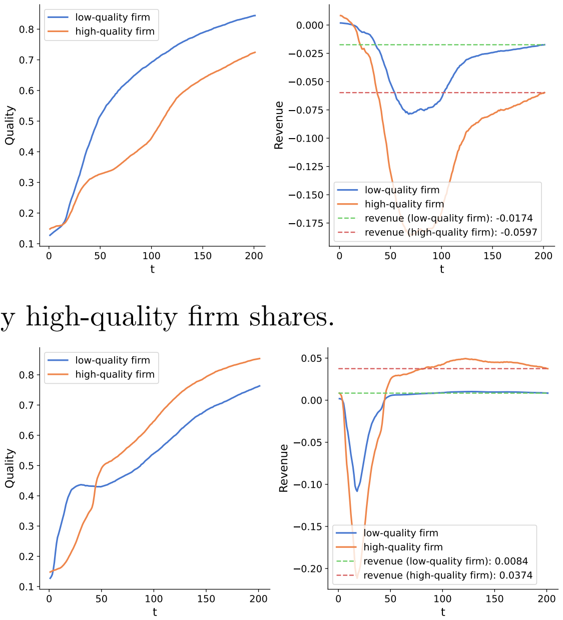
In this section, we introduce our method, Defection-Free Collaborative Learning (Defection-Free CL). Our objectives in designing this algorithm are that

1. for all starting values $(q_{l,0}, q_{h,0})$, neither firm's revenue decreases at any round, and
2. the algorithm converges to the Nash bargaining solution, which we denote (q_l^*, q_h^*). (See Section 3.4.)

The first objective ensures that the algorithm is defection-free. The second seeks a point of convergence that maximizes the joint surplus of the firms. In Section 3.4, we show that Algorithm 5 achieves 1) entirely and achieves 2) for a large range of starting conditions. Before describing our algorithm, we first motivate the Nash bargaining solution as a suitable convergence goal for our problem setting.

Nash Bargaining

In cooperative bargaining, agents determine how to share a surplus amongst themselves. If negotiations fail, each agent is guaranteed some fixed surplus, known as the *disagreement point*. A typical application of bargaining involves deciding how to split a firm's profits amongst its employees. The bargaining framework is suitable for our purposes because the



(a) Only high-quality firm shares.

(b) Only low-quality firm shares.

Figure 3.2: Performance of One-sided Sharing schemes on MNIST. When only the high-quality firm shares, the high-quality firm's revenue becomes negative. When only the low-quality firm shares, both firms have positive, but less, revenue than with our collaboration scheme (Figure 3.3).

firms must agree how to share a "surplus of quality" (i.e. set model qualities relative to each other) so that neither firm's revenue decreases at any one round.

An important framework in cooperative bargaining is Nash bargaining [45], a two-person bargaining scheme, which solves for

$$\begin{aligned} (q_l^*, q_h^*) = \arg\max_{(q_l, q_h)} \quad & N(q_l, q_h, q_{l,0}, q_{h,0}) \\ \text{s.t.} \quad & U_l(q_l, q_h) \geq U_l(q_{l,0}, q_{h,0}) \\ & U_h(q_l, q_h) \geq U_h(q_{l,0}, q_{h,0}), \end{aligned}$$

where

$$N(q_l, q_h, q_{l,0}, q_{h,0}) \overset{\text{def}}{=} (U_l(q_l, q_h) - U_l(q_{l,0}, q_{h,0}))(U_h(q_l, q_h) - U_h(q_{l,0}, q_{h,0})),$$

and $(q_{l,0}, q_{h,0})$ are the initial model qualities of the firms. The *Nash bargaining solution*, (q_l^*, q_h^*), maximizes the product of the *improvement* in the firms' utilities. Therefore, unlike one-sided collaboration, the Nash objective rewards improvement in the low-quality firm's utility as well as the high-quality firm's utility. In Nash bargaining, the *disagreement point*

$(q_{l,0}, q_{h,0})$ determines the surplus for the parties if negotiations fall apart. In our setting, if either firm defects from collaboration, both firms retain their current model qualities. Going forward, we use $N(q_l, q_h)$ as shorthand for $N(q_l, q_h, q_{l,0}, q_{h,0})$. The Nash bargaining solution (q_l^*, q_h^*) has four important properties: 1) it is invariant to affine transformation of the utility functions, 2) it is pareto efficient, 3) it is symmetric, and 4) it is independent of irrelevant alternatives. In fact, the point (q_l, q_h) with these four properties is uniquely the Nash bargaining solution.

The next proposition shows that q_h^* is equivalent to the high-quality firm's maximal quality.

Proposition 2 (Equivalence between maximal quality and the Nash bargaining solution)**.**

$$q_h^* = \max_{x \in \mathcal{X}} q(x).$$

Proof. From Proposition 1, $\frac{\partial U_h}{\partial q_h}$ and $\frac{\partial U_l}{\partial q_h}$ are both non-negative for all $q_l \leq q_h$, and consequently $\frac{\partial N(q_l, q_h)}{\partial q_h} \geq 0$ for all $q_l \leq q_h$. This means that for any q_l, the $N(q_l, q_h)$ can always be improved by increasing q_h. Therefore, q_h^* is necessarily $\max_{x \in \mathcal{X}} q(x)$. □

Section 3.3 shows there's a defection-free scheme in which the low-quality firm shares but the high-quality firm doesn't. In Algorithm 5, we give a way for both firms to contribute to collaboration with neither firm losing revenue at any step. Due to the more equitable design of this collaboration scheme, its dynamics mirror those of Nash bargaining which maximizes the joint surplus of the participants.

The difficulty of designing Algorithm 5 is that, in order to reach (q_l^*, q_h^*) without decreasing revenues at any step, neither firm can improve its quality too much in a given step. Given an increase in the high-quality firm's quality $q_{h,t-1} \rightarrow q_{h,t}$, the low-quality firm can only improve by some limited amount without decreasing the high-firm's revenue (since U_h is decreasing in q_l by Prop. 1). Because of this capped permissible improvement for the low-quality firm, if the high-quality firm converges to q_h^* too quickly, the low-quality firm will never reach q_l^*.

We describe the key steps of Algorithm 5. We also assume that, prior to the algorithm, both firms have saturated training on their own datasets and will only update their models collaboratively going forward. Since U_l and U_h both increase in q_h, the low-quality firm should always share with the high-quality firm. Step 4 ensures this, where the high-quality firm has access to the low-quality firm's loss on its model $x_{h,t-1}$ when updating. As we show in Section 3.4, in order to converge to the Nash bargaining solution, the low-quality firm should not update if $q_{l,t} \geq q_l^*$ or $\rho_{t-1} > \rho^*$. Step 7 ensures this. Since U_h decreases in q_l, the low-quality firm cannot improve its model beyond a certain threshold before the high-quality firm loses revenue. This threshold $\hat{q}_{l,t}$ is computed in Step 8, and in Steps 9-11, the high-quality firm will only collaborate if the collaborative updates to the low-quality firm's model do not improve its quality beyond $\hat{q}_{l,t}$.

Algorithm 5 Defection-Free Collaborative Learning

Input: Low-quality model: $x_{l,0}$. High-quality model: $x_{h,0}$.

Note: We assume both firms are trusted parties and will honestly exchange information. For example, to perform the necessary computations, the high-quality firm requires x_l and $\nabla f(x_h; l)$ from the low-quality firm, and the low-quality firm requires x_h, $\nabla f(x_l; h)$, $f(x_h; h)$, and $f(x_l; h)$ from the high-quality firm.

```
for $t \in [T]$ do
    High-quality Model Update
    Set $\alpha_{h,t} \leq \frac{1}{L}$.
    Update: $x_{h,t} = x_{h,t-1} - \alpha_{h,t} \nabla_{x_{h,t-1}} f_{h,t-1}$.
    Low-quality Model Update
    $x_{l,t} = x_{l,t-1}$.
    if $q_{l,t} < q_l^*$ and $\frac{q_{l,t}}{q_{h,t}} \leq \rho^* = \frac{q_l^*}{q_h^*}$ then
        Compute: $\hat{q}_{l,t} = B\left(\rho_{t-1}, \frac{q_{h,t}}{q_{h,t-1}}\right) q_{h,t}$, where
            $$B(a,b) \overset{\text{def}}{=} 4 - \frac{(4-a)^2}{2(1-a)} \left( b - \sqrt{b^2 - \frac{12(1-a)}{(4-a)^2} b} \right).$$
        while $q_{l,t} \leq \hat{q}_{l,t}$ do
            Set: $\alpha_{l,t}$.
            Update: $x_{l,t} \leftarrow x_{l,t} - \alpha_{l,t} \nabla_{x_{l,t}} f_{l,t}$
Output: $x_{l,T}, x_{h,T}$
```

In the next section we prove the two key properties of Defection-Free Collaborative Learning: 1) it guarantees the firms non-decreasing revenue at every step, and 2) it converges to the Nash bargaining solution for all but trivial starting conditions.

Theory and Analysis

The following proposition shows that Algorithm 5 is defection-free.

Proposition 3 (Non-decreasing revenues)**.** *There exist learning rate schedules $\{\alpha_{l,t}\}_t$ and $\{\alpha_{h,t}\}_t$ such that at no step of Algorithm 5 does either firm's revenue decrease.*

Proof. See Section 3.7. □

We next examine starting conditions for which Algorithm 5 converges to the Nash bargaining solution. Proposition 4 gives a trivial starting condition for which it does not converge.

Proposition 4 (Impossibility of convergence to the Nash bargaining solution)**.** *If $q_{l,0} > q_l^*$, then Algorithm 5 cannot converge to (q_l^*, q_h^*).*

Proof. Since firms do not degrade their model quality, the low-quality firm cannot converge to q_l^*. □

In the next proposition, we show that for all other starting conditions, Algorithm 5 converges to (q_l^*, q_h^*). Our key insight in the proof of this proposition is that if the high-quality firm converges too quickly to q_h^*, the low-quality firm will not be able to make sufficient progress towards q_l^* without violating the no-revenue-loss condition. Therefore, we must design a learning rate schedule for the high-quality firm $\{\alpha_{h,t}\}_t$ such that convergence to q_h^* is properly paced.

Proposition 5 (Convergence to the Nash bargaining solution)**.** *If $q_{l,0} \leq q_l^*$, then there exist learning rate schedules $\{\alpha_{l,t}\}_{t=1}^T$ and $\{\alpha_{h,t}\}_{t=1}^T$ such that after T rounds Algorithm 5 converges to (q_l^*, q_h^*).*

Proof. See Section 3.7. □

Proposition 5 shows that even when both firms myopically attend to improving their own revenues, Algorithm 5 converges to the Nash bargaining solution which maximizes joint surplus. The following theorem gives the rate of convergence to the Nash bargaining solution for convex and L-smooth losses.

Theorem 5 (Convergence Rate of Defection-Free Collaborative Learning)**.** *Suppose $q_{l,0} \leq q_l^*$. Then running Algorithm 5 for T rounds ensures*

$$N(q_l^*, q_h^*) - N(q_{l,T}, q_{h,T}) \lesssim \frac{\|x_{h,0} - x_h^*\|^2}{\sum_{t=1}^T \alpha_{h,t}} + |\rho^* - \rho_T|. \tag{3.5}$$

Proof. See Section 3.7. □

The first term in the bound (3.5) shows that the convergence rate to the Nash bargaining solution is determined by the rate at which q_h converges to q_h^*.

The following corollary shows the rate at which the $|\rho^* - \rho_T|$ term in Theorem 5 vanishes with T.

Corollary 1. *Suppose $q_{l,0} \leq q_l^*$. Running Algorithm 5 for $T \gtrsim \frac{L\|x_{h,0}-x_h^*\|^2}{\epsilon}$ rounds ensures that*

$$N(q_l^*, q_h^*) - N(q_{l,T}, q_{h,T}) \lesssim \frac{\|x_{h,0} - x_h^*\|^2}{\sum_{t=1}^T \alpha_{h,t}} + (4 - 5\rho^*) \log\left(\frac{q_h^*}{q_h^* - \epsilon}\right).$$

Proof. See Section 3.7. □

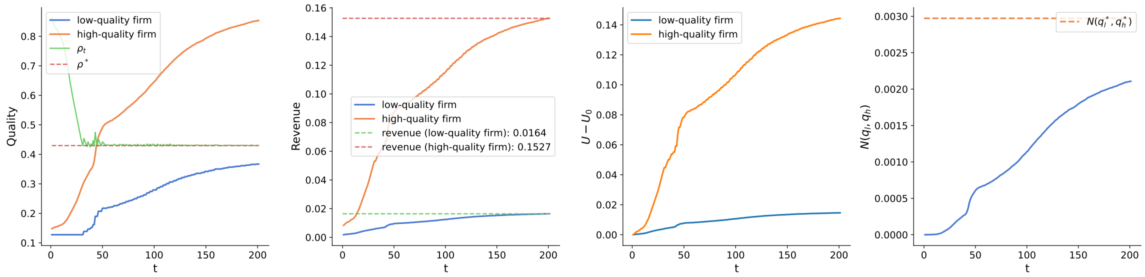


Figure 3.3: Performance of Defection-Free FL on MNIST. Both firms' qualities increase (figure 1), their revenues increase and approach a higher level than under One-sided Collaboration (figure 2), and the firms' qualities approach the Nash bargaining solution (figure 4).

3.5 Experiments

All algorithms in our are implemented with PyTorch [47]. Our general experimental setup is the following. We construct three datasets: the low-quality firm's dataset $\mathcal{D}_{l,\text{train}}$, the high-quality firm's dataset $\mathcal{D}_{h,\text{train}}$, and a common test set for both firms $\mathcal{D}_{\text{target}}$. The datasets are constructed such that $\mathcal{D}_{l,\text{train}} \not\sim \mathcal{D}_{\text{target}}$ and $\mathcal{D}_{h,\text{train}} \not\sim \mathcal{D}_{\text{target}}$, but $\mathcal{D}_{l,\text{train}} \cup \mathcal{D}_{h,\text{train}} \sim \mathcal{D}_{\text{target}}$, i.e. neither firm's individual distribution matches the target distribution, but their combined datasets are distributed as the target distribution, incentivizing them to share. We use cross-entropy loss and PyTorch's built-in SGD optimizer for all experiments. Code for all experiments is available at this github repo.

MNIST We use a LeNet-5 model [48], set $|\mathcal{D}_{l,\text{train}}| = |\mathcal{D}_{h,\text{train}}| = 1000$, and use the MNIST test set as $\mathcal{D}_{\text{target}}$. We construct $\mathcal{D}_{l,\text{train}}$ so that $\hat{F}(5) = 0.8$ and $\mathcal{D}_{h,\text{train}}$ so that $\hat{F}(5) = 0.2$, where $\hat{F}$ is the empirical CDF over the label space. We train the high-quality firm's model for 10 initial epochs, and for all models and experiments set the learning rate to 0.01.

Defection-Free Collaborative Learning (Fig. 3.3). Since the low-quality firm shares with the high-quality firm, the high-quality firm improves maximally. The high-quality firm only shares with the low-quality firm to the extent that neither firm's revenue decreases. Under this sharing scheme, we see in the first figure that both firms' qualities increase, and the ratio of their qualities converges to the optimal ratio. The second figure shows that revenues increase (do not decrease), and notably their revenues reach a higher level than under One-sided Collaboration (Section 3.3). Finally, the last figure shows that the Nash bargaining objective approaches its maximal value, showing convergence to the Nash bargaining solution.

3.6 Conclusion

We introduce a defection-free collaborative learning scheme in which participants iteratively optimize their models by sharing training resources, without losing utility at any round and having cause to defect from participation. Framing the collaborative learning system as a duopoly of competitive firms, we show that both firms can improve their model qualities by sharing data with each other without losing revenue at any round. We describe other collaboration schemes for which this is not possible. Notably, even when both firms myopically focus on improving their own revenues, we show that our algorithm converges to the Nash bargaining solution, thus optimizing for joint surplus.

3.7 Proofs of Theoretical Results

Proofs for Section 3.2

Proof of Lemma 21. Let $\hat{\theta}_l$ be the type of the consumer who is indifferent between buying from the low-quality firm and not buying at all. Then, based on the consumer's utility function (3.2),

$$\hat{\theta}_l q_l - p_l = 0. \tag{3.6}$$

Let $\hat{\theta}_h$ be the type of the consumer who is indifferent between buying from the high-quality firm and low-quality firm. Then, from (3.2),

$$\hat{\theta}_h q_l - p_l = \hat{\theta}_h q_h - p_h. \tag{3.7}$$

Therefore any consumer with type $\theta \in [\hat{\theta}_l, \hat{\theta}_h)$ will buy from the low-quality firm and any consumer with type $\theta \in [\hat{\theta}_h, 1]$ will buy from the high-quality firm, giving demands $D_l = \hat{\theta}_h - \hat{\theta}_l$ and $D_h = 1 - \hat{\theta}_h$. Solving (3.6) and (3.7) for $\hat{\theta}_l$ and $\hat{\theta}_h$ completes the proof. □

Proof of Lemma 22. From Lemma 21, the demand for the low-quality firm is $D_l = \frac{p_h - p_l}{q_h - q_l} - \frac{p_l}{q_l}$, yielding low-quality firm utility

$$U_l = p_l\left(\frac{p_h - p_l}{q_h - q_l} - \frac{p_l}{q_l}\right). \tag{3.8}$$

To maximize its utility, the low-quality firm sets price

$$\begin{aligned} p_l^* &= \arg\max_{p_l} \frac{\partial U_l}{\partial p_l} \\ &= \arg\max_{p_l} \left(\frac{p_h - 2p_l}{q_h - q_l} - \frac{2p_l}{q_l}\right) \end{aligned}$$

$$= \frac{q_l p_h}{2q_h}. \tag{3.9}$$

Similarly, demand for the high-quality firm is $D_h = 1 - \frac{p_h - p_l}{q_h - q_l}$, yielding high-quality firm utility

$$U_h = p_h\left(1 - \frac{p_h - p_l}{q_h - q_l}\right). \tag{3.10}$$

To maximize its utility, the high-quality firm sets price

$$\begin{aligned} p_h^* &= \arg\max_{p_h} \frac{\partial U_h}{\partial p_h} \\ &= \arg\max_{p_h} \left(1 - \frac{2p_h - p_l}{q_h - q_l}\right) \\ &= \frac{p_l + (q_h - q_l)}{2}. \end{aligned} \tag{3.11}$$

Resolving (3.9) and (3.11) yields

$$p_l^* = \frac{q_l(q_h - q_l)}{4q_h - q_l} \tag{3.12}$$

and

$$p_h^* = \frac{2q_h(q_h - q_l)}{4q_h - q_l}. \tag{3.13}$$

Finally, evaluating (3.8) and (3.10) at the optimal prices (3.12) and (3.13) yields the price-optimal utilities (3.4). □

Proof of Proposition 1. The proposition follows from observing the partial derivatives of the firms' utility functions. For $q_l \leq q_h$,

$$\frac{\partial U_h}{\partial q_h} = \frac{4q_h(4q_h^2 - 3q_h q_l + 2q_l^2)}{(4q_h - q_l)^3} \geq 0,$$

$$\frac{\partial U_l}{\partial q_h} = \frac{q_l^2(2q_h + q_l)}{(4q_h - q_l)^3} \geq 0,$$

$$\frac{\partial U_l}{\partial q_l} = \frac{q_h^2(4q_h - 7q_l)}{(4q_h - q_l)^3} \begin{cases} \geq 0 & \text{if } q_l \leq \frac{4}{7}q_h \\ < 0 & \text{if } q_l > \frac{4}{7}q_h \end{cases}$$

and

$$\frac{\partial U_h}{\partial q_l} = -\frac{4q_h^2(2q_h + q_l)}{(4q_h - q_l)^3} \leq 0.$$

Figure 3.4 provides a graphical representation of this proposition. □

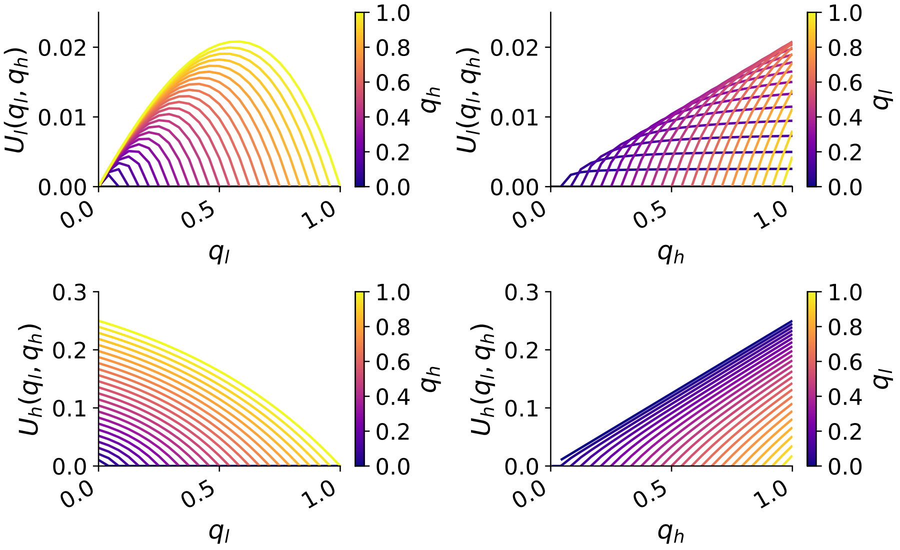


Figure 3.4: This figure shows how the firms' utilities vary with model quality. U_l and U_h are both increasing in q_h, U_h is decreasing in q_l, and U_l is increasing in q_l for $q_l \leq \frac{4q_h}{7}$ and decreasing in q_l otherwise.

Proofs for Section 3.4

Proof of Proposition 3. Suppose that at round t, given current qualities $q_{l,t-1}$ and $q_{h,t-1}$, the high-quality firm improves to $q_{h,t}$. Then, in order for neither firm to lose revenue, $q_{l,t}$ must be such that

$$\frac{4q_{h,t}^2(q_{h,t}-q_{l,t})}{(4q_{h,t}-q_{l,t})^2} \geq \frac{4q_{h,t-1}^2(q_{h,t-1}-q_{l,t-1})}{(4q_{h,t-1}-q_{l,t-1})^2} \tag{3.14}$$

and

$$\frac{q_{l,t}q_{h,t}(q_{h,t}-q_{l,t})}{(4q_{h,t}-q_{l,t})^2} \geq \frac{q_{l,t-1}q_{h,t-1}(q_{h,t-1}-q_{l,t-1})}{(4q_{h,t-1}-q_{l,t-1})^2}. \tag{3.15}$$

Rearranging terms, (3.14) can be written as an inequality involving a convex quadratic of $q_{l,t}$:

$$\begin{aligned}
&[4q_{h,t-1}^2(q_{h,t-1}-q_{l,t-1})]q_{l,t}^2 \\
&\quad + [4(4q_{h,t-1}-q_{l,t-1})^2 q_{h,t}^2 - 32q_{h,t-1}^2(q_{h,t-1}-q_{l,t-1})q_{h,t}]q_{l,t} \\
&\quad + [64q_{h,t-1}^2(q_{h,t-1}-q_{l,t-1})q_{h,t}^2 - 4(4q_{h,t-1}-q_{l,t-1})^2 q_{h,t}^3] < 0.
\end{aligned}$$

The right-most root of this quadratic is

$$q_{l,t}^h = 4q_{h,t} - \frac{(4-\rho_{t-1})^2}{2(1-\rho_{t-1})}\left(\frac{q_{h,t}^2}{q_{h,t-1}} - \sqrt{\frac{q_{h,t}^4}{q_{h,t-1}^2} - \frac{12(1-\rho_{t-1})}{(4-\rho_{t-1})^2}\frac{q_{h,t}^3}{q_{h,t-1}}}\right).$$

Similarly, (3.15) can be written as an inequality involving a convex quadratic of $q_{l,t}$:

$$\begin{aligned}&[q_{l,t-1}q_{h,t-1}(q_{h,t-1}-q_{l,t-1})+(4q_{h,t-1}-q_{l,t-1})^2q_{h,t}]q_{l,t}^2\\&\quad+[-8q_{l,t-1}q_{h,t-1}(q_{h,t-1}-q_{l,t-1})q_{h,t}-(4q_{h,t-1}-q_{l,t-1})^2q_{h,t}^2]q_{l,t}\\&\quad+[16q_{l,t-1}q_{h,t-1}(q_{h,t-1}-q_{l,t-1})q_{h,t}^2]<0.\end{aligned}$$

The right-most root of this quadratic is

$$\begin{aligned}q_{l,t}^l=&\left(\frac{1}{2((1-\rho_{t-1})\rho_{t-1}q_{h,t-1}+(4-\rho_{t-1})^2q_{h,t})}\right)\cdot\\&\left(8(1-\rho_{t-1})\rho_{t-1}q_{h,t-1}+(4-\rho_{t-1})^2q_{h,t}\right.\\&\quad\left.+(4-\rho_{t-1})\sqrt{(4-\rho_{t-1})^2q_{h,t}^2-48\rho_{t-1}(1-\rho_{t-1})q_{h,t-1}q_{h,t}}\right).\end{aligned}$$

It can be verified with graphing software that for all feasible parameters, $q_{l,t}^h\le q_{l,t}^l$. Therefore, the low-quality firm can improve its quality to at most

$$\hat{q}_{l,t}=4q_{h,t}-\frac{(4-\rho_{t-1})^2}{2(1-\rho_{t-1})}\left(\frac{q_{h,t}^2}{q_{h,t-1}}-\sqrt{\frac{q_{h,t}^4}{q_{h,t-1}^2}-\frac{12(1-\rho_{t-1})}{(4-\rho_{t-1})^2}\frac{q_{h,t}^3}{q_{h,t-1}}}\right),$$

before at least one of the firms loses revenue. Algorithm 5 ensures that $q_{l,t}$ does not exceed $\hat{q}_{l,t}$.

It remains to prove that there exist learning rate sequences $\{\alpha_{l,t}\}_t$ and $\{\alpha_{h,t}\}_t$ that respect the constraint $q_{l,t}\le\hat{q}_{l,t}$. Since improvement in q_h increases the revenues of both firms (Prop. 1), the high-quality firm can set any learning rate schedule $\{\alpha_{h,t}\}_t$ without violating the no-revenue-loss constraints (3.14) and 3.15. For the low-quality firm's learning rate schedule, note that $f_l(x)$, as the average of convex functions $f(x;l)$ and $f(x;h)$, is also convex. Therefore,

$$\begin{aligned}f_{l,t}&\ge f_{l,t-1}+\nabla_{x_{l,t-1}}f_{l,t-1}^T(x_{l,t}-x_{l,t-1})\\&=f_{l,t-1}-\alpha_{l,t}\|\nabla_{x_{l,t-1}}f_{l,t-1}\|^2.\end{aligned}$$

Rearranging terms,

$$\begin{aligned}\alpha_{l,t}&\ge\frac{f_{l,t-1}-f_{l,t}}{\|\nabla_{x_{l,t-1}}f_{l,t-1}\|^2}\\&=\frac{q_{l,t}-q_{l,t-1}}{\|\nabla_{x_{l,t-1}}f_{l,t-1}\|^2}.\end{aligned}$$

Therefore, setting $\alpha_{l,t}=\min\left\{\frac{\hat{q}_{l,t}-q_{l,t-1}}{\|\nabla_{x_{l,t-1}}f_{l,t-1}\|^2},1\right\}$ ensures that the low-quality firm's updated quality $q_{l,t}$ does not exceed $\hat{q}_{l,t}$. □

Proof of Proposition 5. We handle the proof in cases.

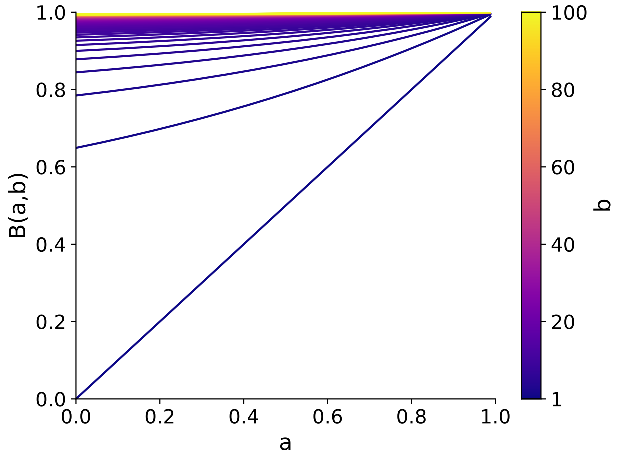


Figure 3.5: $B(a, b) \geq a$ for all $b \geq 1$.

Case 1: $q_{l,0} \leq q_l^*$ and $\rho_0 \geq \rho^*$.

When $\frac{q_{l,t-1}}{q_{h,t}} \geq \rho^*$, the low-quality firm does not update (line 7 of Alg. 5). Once the high-quality firm improves sufficiently so that $\frac{q_{l,t}}{q_{h,t}} = \rho^*$ (note that such a t exists if $q_{l,0} \leq q_l^*$), then convergence is guaranteed. To see this, we use the following lemma.

Lemma 24. *$B(a, b) \geq a$ for all $b \geq 1$. (See Figure 3.5 for pictorial proof.)*

Consider step $t+1$ at which $\rho_t = \frac{q_{l,t}}{q_{h,t}} = \rho^*$. Given the high-quality firm's improvement $q_{h,t} \to q_{h,t+1}$, if the low-quality firm improves to $q_{l,t+1} = \hat{q}_{l,t+1}$, by Lemma 24, $\rho_{t+1} \geq \rho_t$. Therefore the low-quality firm can always improve to some level $q_{l,t+1} \in [q_{l,t}, \hat{q}_{l,t+1}]$ and ensure that $\rho_{t+1} = \rho^*$ with neither firm losing revenue. Maintaining this improvement schedule, once the high-quality firm improves to q_h^* (using any sequence of learning rates $\{\alpha_{h,t}\}_t$), the low-quality firm will be able to reach q_l^* by observing the constraint in lines 9-11 of Alg. 5.

Case 2: $q_{l,0} \leq q_l^*$ and $\rho_0 < \rho^*$.

Our strategy for this case will be to show there exist sequences of learning rates $\{\alpha_{h,t}\}_t$ and $\{\alpha_{l,t}\}_t$ such that $\sum_{t=1}^{T}(\rho_t - \rho_{t-1}) = \rho_T - \rho_0 \geq \rho^* - \rho_0$. We will do this by lower-bounding the quality-ratio gaps $\rho_t - \rho_{t-1} = B(\rho_{t-1}, {q_{h,t}}/{q_{h,t-1}}) - \rho_{t-1}$.

For each $\rho \leq 1$, there is a point (possibly infinite)

$$b_\rho \overset{\text{def}}{=} \max\{b \geq 1 : (4 - 5\rho)\log_{10} b \leq B(\rho, b) - \rho\}.$$

That is, for a given ρ, b_ρ is the point at which $(4 - 5\rho)\log b$ goes from being a lower to an

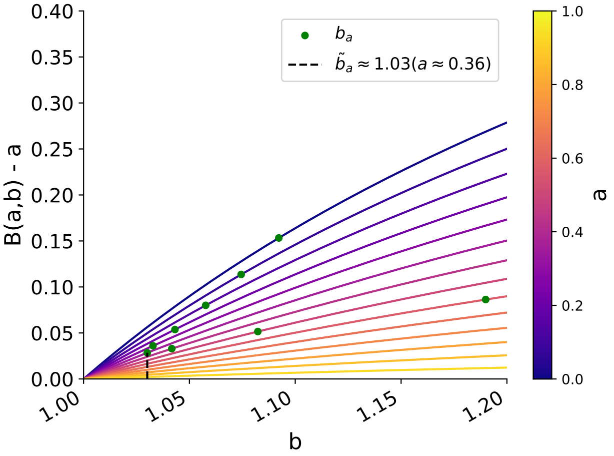


Figure 3.6: The green dots indicate, for a given ${}^{q_{l,t-1}}\!/_{q_{h,t-1}}$ (symbolized by a), the upper bound on ${}^{q_{h,t}}\!/_{q_{h,t-1}}$ that ensures convergence to the Nash bargaining solution.

upper bound on $B(\rho, b) - \rho$. Define $\tilde{b}$ as the smallest such point over all $\rho \leq 1$, so

$$\tilde{b} \overset{\text{def}}{=} \min_{\rho \leq 1} b_\rho.$$

Figure 3.6 plots b_ρ for various values of ρ and shows that $\tilde{b} \approx 1.03 = b_{\rho \approx 0.33}$.

By definition of $\tilde{b}$, $(4 - 5\rho)\log_{10} b \leq B(\rho, b) - \rho$ for any $\rho \leq 1$ and $b \leq \tilde{b}$. Suppose the high-quality firm maintains a learning rate schedule $\{\alpha_{h,t}\}_t$ such that ${}^{q_{h,t}}\!/_{q_{h,t-1}} \leq \tilde{b}$ for all t and T is such that $q_h^* - q_{h,T} \leq \epsilon$. Then

$$\begin{aligned}\sum_{t=1}^{T}(\rho_t - \rho_{t-1}) &= \sum_{t=1}^{T}(B(\rho_{t-1}, {}^{q_{h,t}}\!/_{q_{h,t-1}}) - \rho_{t-1}) \\ &\overset{(i)}{\geq} \sum_{t=1}^{T}(4 - 5\rho_{t-1})\log_{10}({}^{q_{h,t}}\!/_{q_{h,t-1}}) \\ &\overset{(ii)}{\geq} (4 - 5\rho^*)\log_{10}({}^{q_{h,T}}\!/_{q_{h,0}}) \\ &\geq (4 - 5\rho^*)\log_{10}({}^{(q_h^*-\epsilon)}\!/_{q_{h,0}}),\end{aligned}$$

where (i) is due to ${}^{q_{h,t}}\!/_{q_{h,t-1}} \leq \tilde{b}$, and (ii) is due to the fact that $\rho_0 \leq \rho^*$ and Lemma 24.

Figure 3.7 shows that $(4 - 5\rho^*)\log_{10}({}^{q_h^*}\!/_{q_{h,0}}) \geq \rho^* - \rho_0$, so

$$(4 - 5\rho^*)\log_{10}\left(\frac{q_h^* - \epsilon}{q_{h,0}}\right) = (4 - 5\rho^*)\left(\log_{10}\left(\frac{q_h^*}{q_{h,0}}\right) - \log_{10}\left(\frac{q_h^*}{q_h^* - \epsilon}\right)\right)$$

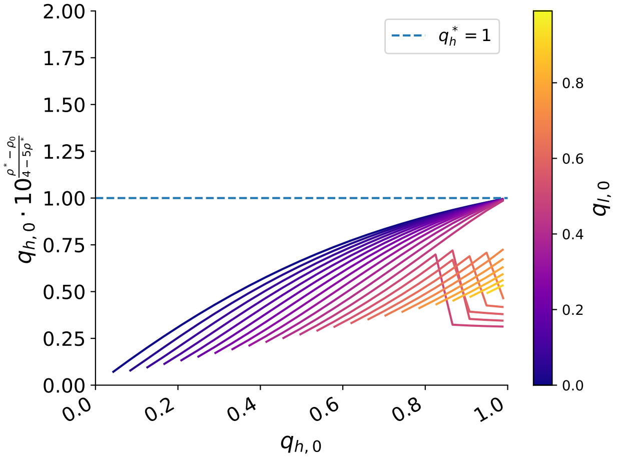


Figure 3.7: Empirical verification of the inequality: $(4 - 5\rho^*) \log_{10}(q_h^*/q_{h,0}) \geq \rho^* - \rho_0$

$$\geq (\rho^* - \rho_0) - (4 - 5\rho^*) \log_{10}\left(\frac{q_h^*}{q_h^* - \epsilon}\right).$$

Therefore $\rho^* - \rho_T \leq (4 - 5\rho^*) \log_{10}\left(\frac{q_h^*}{q_h^* - \epsilon}\right)$.

It remains to show that there exists a sequence of learning rates $\{\alpha_{h,t}\}_t$ such that $q_{h,t}/q_{h,t-1} \leq \tilde{b}$, and T such that $q_h^* - q_{h,T} \leq \epsilon$. Let $\alpha_{h,t} = \min\left\{\frac{(\tilde{b}-1)q_{h,t-1}}{\|\nabla_{x_{h,t-1}} f_{h,t-1}\|^2}, \frac{1}{L}\right\}$. We analyze what happens when $\alpha_{h,t}$ is each of the values in the *min* expression.

First, suppose $\alpha_{h,t} = \frac{(\tilde{b}-1)q_{h,t-1}}{\|\nabla_{x_{h,t-1}} f_{h,t-1}\|^2}$ for all t. f_h, as the average of L-smooth and convex functions, is also L-smooth and convex, so that

$$\frac{q_{h,t-1} + \frac{\alpha_{h,t}}{2}\|\nabla_{x_{h,t-1}} f_{h,t-1}\|^2}{q_{h,t-1}} \leq \frac{q_{h,t}}{q_{h,t-1}} \leq \frac{q_{h,t-1} + \alpha_{h,t}\|\nabla_{x_{h,t-1}} f_{h,t-1}\|^2}{q_{h,t-1}}.$$

Therefore, the choice of $\alpha_{h,t}$ guarantees that $\frac{\tilde{b}+1}{2} \leq \frac{q_{h,t}}{q_{h,t-1}} \leq \tilde{b}$, giving $\frac{q_{h,T}}{q_{h,0}} \geq \left(\frac{\tilde{b}+1}{2}\right)^T$. From this we see that setting $T \geq \frac{\log(q_h^*/q_{h,0})}{\log((\tilde{b}+1)/2)}$ guarantees convergence to q_h^* in T' steps.

Now suppose $\alpha_{h,t} = \frac{1}{L}$ for all t. Under this condition, standard convergence analysis for gradient descent on convex and L-smooth functions gives

$$f_{h,T} - f_h^* \leq \frac{L\|x_{h,0} - x_h^*\|^2}{2T}.$$

Therefore, $f_{h,T} - f_h^* \leq \epsilon$ after $T = \frac{L\|x_{h,0} - x_h^*\|^2}{2\epsilon}$ rounds.

From the above analysis, we see that after at most $T = \frac{\log(q_h^*/q_{h,0})}{\log((\tilde{b}+1)/2)} + \frac{L\|x_{h,0}-x_h^*\|^2}{2\epsilon}$ rounds, $f_{h,T} - f_h^* = q_h^* - q_{h,T} \leq \epsilon$, completing the proof. □

Proof of Theorem 5. By Taylor's theorem,

$$\begin{aligned}
N(q_l^*, q_h^*) &\leq N(q_{l,T}, q_{h,T}) + \frac{\partial N(q_l, q_h)}{\partial q_l}(q_l^* - q_{l,T}) + \frac{\partial N(q_l, q_h)}{\partial q_h}(q_h^* - q_{h,T}) \\
&\quad + \left(\max_{q_l, q_h} \frac{\partial^2 N(q_l, q_h)}{\partial q_l^2}\right)\frac{(q_l^* - q_{l,T})^2}{2} + \left(\max_{q_l, q_h} \frac{\partial^2 N(q_l, q_h)}{\partial q_h^2}\right)\frac{(q_h^* - q_{h,T})^2}{2} \\
&\quad + \left(\max_{q_l, q_h} \frac{\partial^2 N(q_l, q_h)}{\partial q_h \partial q_l}\right)(q_l^* - q_{l,T})(q_h^* - q_{h,T}) \\
&\overset{(i)}{\leq} c_1(q_h^* - q_{h,T}) + c_2(\rho^*(q_h^* - q_{h,T}) + q_{h,T}|\rho^* - \rho_T|) \\
&\lesssim (q_h^* - q_{h,T}) + |\rho^* - \rho_T|,
\end{aligned}$$

where (i) follows from the fact that the gradients of N are bounded by small constants (can be verified with graphing software), qualities $q \in [0, 1]$, and $q_l^* - q_{l,T} = \rho^* q_h^* - \rho_T q_{h,T} \leq \rho^*(q_h^* - q_{h,T}) + q_{h,T}|\rho^* - \rho_T|$.

We now bound $q_h^* - q_{h,T}$. Note that f_h, as the average of L-smooth and convex functions, is also L-smooth and convex. Therefore,

$$\begin{aligned}
f_{h,t} &\overset{(i)}{\leq} f_{h,t-1} + \left(-\alpha_{h,t} + \frac{L\alpha_{h,t}^2}{2}\right)\|\nabla_{x_{h,t-1}} f_{h,t-1}\|^2 \\
&\overset{(ii)}{\leq} f_{h,t-1} - \frac{\alpha_{h,t}}{2}\|\nabla_{x_{h,t-1}} f_{h,t-1}\|^2 \\
&\overset{(iii)}{\leq} f_h^* + \nabla_{x_{h,t-1}} f_{h,t-1}^T (x_{h,t-1} - x_h^*) - \frac{\alpha_{h,t}}{2}\|\nabla_{x_{h,t-1}} f_{h,t-1}\|^2 \\
&= f_h^* + \frac{2}{\alpha_{h,t}}(\|x_{h,t-1} - x_h^*\|^2 - \|x_{h,t} - x_h^*\|^2),
\end{aligned}$$

where (i) is due to L-smoothness of f_h, (ii) is due to $\alpha_{h,t} \leq \frac{1}{L}$, and (iii) is due to convexity of f_h. Rearranging terms and summing over t,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\alpha_{h,t}}{2}(f_{h,t} - f_h^*) &\leq \sum_{t=1}^{T} \|x_{h,t-1} - x_h^*\|^2 - \|x_{h,t} - x_h^*\|^2 \\
&\leq \|x_{h,0} - x_h^*\|^2. \qquad (3.16)
\end{aligned}$$

Since $\{f_{h,t}\}_t$ are decreasing, (3.16) implies that

$$f_{h,T} - f_h^* \leq \frac{2\|x_{h,0} - x_h^*\|^2}{\sum_{t=1}^{T} \alpha_{h,t}}.$$

Noting that $f_{h,T} - f_h^* = q_h^* - q_{h,T}$ completes the proof. □

Proof of Corollary 1. Due to Theorem 5, showing that $|\rho^* - \rho_T| \leq (4 - 5\rho^*) \log\left(\frac{q_h^*}{q_h^* - \epsilon}\right)$ if $T \gtrsim \frac{L\|x_{h,0} - x_h^*\|^2}{\epsilon}$ completes the proof. We handle it in the same cases as in the proof of Proposition 5.

Case 1: $\rho_0 \geq \rho^*$. From lines 9-11 of Algorithm 5, the low-quality firm will not update its model until after round T, where $\rho_T = \rho^*$. With only the high-quality firm updating before this point, the firms' qualities will have reached a ratio ρ^* by T steps if $\frac{q_{l,0}}{q_{h,T}} = \rho^*$. Dividing both sides of this equation by $q_{h,0}$ and rearranging terms, $\frac{q_{h,T}}{q_{h,0}} = \frac{\rho_0}{\rho^*}$. As we showed for this case in the proof of Proposition 5, $\frac{q_{h,t}}{q_{h,t-1}} \leq \tilde{b}$. Therefore,

$$\frac{q_{h,T}}{q_{h,0}} = \frac{\rho_0}{\rho^*} \leq \tilde{b}^T,$$

which gives $T \geq \frac{\log(\rho_0/\rho^*)}{\log(\tilde{b})}$. That is, after $\frac{\log(\rho_0/\rho^*)}{\log(\tilde{b})}$ steps, $\rho_T = \rho^*$. As discussed in the proof of Proposition 5, the firms can maintain a quality ration of ρ^* for all future rounds, making $|\rho^* - \rho_T| = 0$.

Case 2: $\rho_0 < \rho^*$. As the proof of this case in Proposition 5 directly shows, $\rho^* - \rho_T \leq (4 - 5\rho^*) \log\left(\frac{q_h^*}{q_h^* - \epsilon}\right)$ if $T \geq \frac{\log(q_h^*/q_{h,0})}{\log((\tilde{b}+1)/2)} + \frac{L\|x_{h,0} - x_h^*\|^2}{2\epsilon}$.

Combining Cases 1 and 2, if $T \geq \max\left\{\frac{\log(\rho_0/\rho^*)}{\log(\tilde{b})}, \frac{\log(q_h^*/q_{h,0})}{\log((\tilde{b}+1)/2)} + \frac{L\|x_{h,0} - x_h^*\|^2}{2\epsilon}\right\}$, then $|\rho^* - \rho_T| \leq (4 - 5\rho^*) \log\left(\frac{q_h^*}{q_h^* - \epsilon}\right)$, which completes the proof. □

The following lemma further characterizes the Nash bargaining solution in our problem setting.

Lemma 25. *For all ρ_0 s.t. $\rho_0 \leq \rho^*$, $\rho^* \leq 0.43$.*

Proof of Lemma 25. The Nash bargaining objective evaluated at $q_h^* = 1$ is

$$N(q_l, q_h^*) = \left(\frac{q_l(1 - q_l)}{(4 - q_l)^2} - U_{l,0}\right)\left(\frac{4(1 - q_l)}{(4 - q_l)^2} - U_{h,0}\right), \tag{3.17}$$

where $U_{h,0} \overset{\text{def}}{=} U_h(q_{l,0}, q_{h,0})$ and $U_{l,0} \overset{\text{def}}{=} U_l(q_{l,0}, q_{h,0})$.

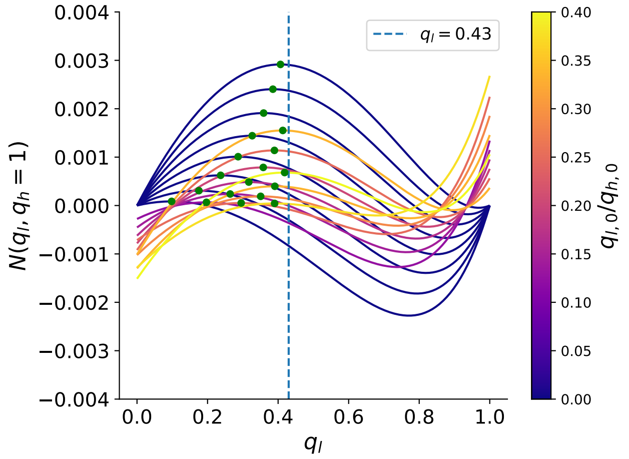


Figure 3.8: For a range of initial qualities and $q_h = q_h^* = 1$, the green dots mark the Nash bargaining solution. The x-values of these points are smaller than 0.43.

Differentiating (3.17) with respect to q_l,

$$\frac{\partial N(q_l, q_h^*)}{\partial q_l} = \left(\frac{1}{(4-q_l)^5}\right) \cdot \Big((7U_{h,0} + U_{h,0}\rho_0 + 4)q_l^3 + (-60U_{h,0} - 6U_{h,0}\rho_0 + 32)q_l^2 + (144U_{h,0} - 52)q_l + (-64U_{h,0} + 32U_{h,0}\rho_0 + 16)\Big). \tag{3.18}$$

The roots of (3.18) correspond to the roots of the cubic numerator. It can be verified with graphing software that over all starting points $(q_{l,0}, q_{h,0})$ such that $\rho_0 \leq \rho^*$, the roots q_l^* of this cubic are at most 0.43. (See Figure 3.8 for empirical evidence.) □

Chapter 4

Privacy Dynamics in Systems of Learning Agents

This chapter is based on Ananthakrishnan et al. 2024, published in *Symposium on the Foundations of Responsible Computing (2024)*

4.1 Introduction

The question of how to define and preserve privacy in the age of machine learning has been a topic of ongoing debate in the computer science and policy communities [49]. The widely accepted theoretical framework of differential privacy [50] formalizes privacy as the ability to withstand membership inference attacks. That is, differential privacy ensures that the output of a computation obfuscates whether a particular data point was present in the input.

However, the practical implementations of differential privacy has been fraught with challenges. There has been significant debate around how to interpret the key privacy parameter ε and how to choose it [51]. This is especially true when data is continuously collected from users (what does it mean to have a guarantee of $\varepsilon = 1$ *per data point* when a user's data is continuously collected?) This has also led to controversies where companies have claimed their algorithms are private, when in fact the chosen ε value confers negligible protection [52]. Further complicating matters, there are multiple variants and extensions of differential privacy—e.g. (ε, δ)-DP [50], Reyni-DP [53], Gaussian-DP [54], etc.—each with different parameters and interpretations.

Perhaps more fundamentally, a growing body of work argues that the public's understanding of privacy is drastically different from differential privacy [55, 56]. While differential privacy focuses on membership inference, privacy is more commonly understood to mean the prevention of the platform using one's data in ways that are misaligned with the individual's interests, such as price discrimination or other exploitative practices.

This work seeks to provide a new perspective on privacy that bridges the gap between the theoretical computer science view and the public's intuitive understanding. We develop a game-theoretic model of privacy that allows us to analyze the effect of privacy choices on all the stakeholders. Additionally, the framework shows how to derive *optimal* privacy mechanisms that balance the gain in privacy with loss of accuracy in order to maximize net utility. In our model, a "principal" (e.g., a platform or seller) can observe signals from "agents" (e.g., users or buyers) and use this information to maximize its own profit, while the agents have an incentive to obfuscate their data to prevent exploitation. We focus on a price-discrimination setting involving interactions between buyers and sellers.

We show that "buyer-induced privacy" behavior, which resembles randomized response, arises endogenously as an equilibrium strategy. Furthermore, we find that the seller is often better off *committing* to not observing the agents' data at all ("seller-induced privacy"), as the revenue loss from buyer-induced privacy can be substantial. Finally, we extend our analysis to a dynamic setting where the seller is a learning agent who interacts with multiple buyers over time. We demonstrate how a simple external auditing mechanism can implement the sellers's commitment to privacy and lead to an equilibrium with endogenously arising privacy-preserving behavior.

Our results provide a new framework for understanding privacy that encompasses both the theoretical guarantees of differential privacy and the practical, user-centric notion of privacy. By modeling privacy as an emergent property of an economic system, we hope to offer insights that can inform the design of privacy-preserving platforms and policies.

Motivating example. In the absence of regulation, online retailers may price discriminate based on information they have collected about past purchases of the customers. Some customers may be willing to pay more for a good than others, perhaps due to innate preferences for certain types of good or because they have more disposable income. The retailer wants to identify customers with higher valuations and charge them higher prices in order to maximize their revenue.

Since customers are aware of the potential for price discrimination, they may engage in evasive action to protect their privacy. Customers may avoid choosing goods that signal their true preferences for less consequential purchases, e.g., a high-income customer choosing between an expensive water bottle that is slightly better than a cheaper option may opt to buy the cheaper bottle in an attempt to obscure their income status. This evasive action imposes a cost on the customer, who misses out on buying their truly preferred product, and also on the retailer, who would have preferred to sell the more expensive product.

What are the behaviors that arise at equilibrium? What if the seller can credibly commit to not price discriminate? How do these behaviors change in more realistic settings where game parameters are not known and strategies must be learned based on past interactions? These are questions we answer in this paper.

Preview of contributions

We introduce a price-discrimination game in Definition 4 that involves buyers of two types—one with a high valuation and one with a low valuation of an item. A seller may potentially track buyers' signals that reveal their valuations. We characterize the perfect Bayes Nash equilibrium of this game in Theorem 6 and show that a buyer-induced privacy mechanism emerges in the equilibrium. That is, the buyer with a high valuation, with some probability, chooses an evasive action to appear to have a low valuation.

We then introduce commitment ability for the seller wherein a seller can commit to not track buyers' signals with some probability. In the price-discrimination game with commitment, the equilibrium response (Corollary 3) results in seller-induced privacy, which obviates the need for buyer-induced privacy. That is, with some probability, the seller chooses to commit to respect privacy and voluntarily does not track signals. Due to this privacy commitment from the seller, it is optimal for buyers to truthfully report their type. We call this seller-induced privacy the "commitment strategy" and denote the resulting utility $\mathbb{U}_1^*$.

In Section 4.3, we remove the seller's commitment ability but give buyers access to the seller's historical pricing. We model this as a repeated interaction between a seller and buyers with each buyer participating in only one round. The pricing history is used by buyers to construct the seller's "reputation" (i.e., an estimate of the probability of price discrimination), which buyers then use to inform their signaling strategy. We model the buyers as using a reputation construction procedure that satisfies a consistency condition given in Definition 6, which requires that the reputation is able to differentiate between sellers employing price-discriminating strategies and non-price-discriminating strategies. In Proposition 6, we show the existence of such a reputation mechanism using the available history. We show that consistent reputation can yield seller-induced privacy (i.e., ignoring signals), depending on the model of the seller; we consider no-regret and no-policy-regret sellers. Our findings are:

1. With a no-regret seller, there could be no seller-induced privacy. That is, the seller can use signals and price discriminate in every round and still be no-regret (Proposition 7).

2. Regret minimization achieves strictly less average utility (asymptotically) than $\mathbb{U}_1^*$ (Proposition 8).

3. Employing the commitment strategy in every round is a no-policy-regret algorithm for the seller (Proposition 9).

4. Employing the commitment strategy in every round ensures the seller (asymptotically) an average utility of $\mathbb{U}_1^*$. This the highest possible average utility achievable (asymptotically) in the repeated interaction (Proposition 10).

Related work

Our work sits at the intersection of many areas, ranging from classical economics to online learning.

There is a vast literature on *privacy* in computer science studying mechanisms for notions of privacy such as differential privacy [50]. The mechanisms arising in our setting resemble mechanisms in these works. We observe local privacy (buyer-induced privacy) where users add noise to their data. We also observe central privacy (seller-induced privacy) where the platform ensures similar outcomes for different user data.

Literature in economics studies the economic implications of enacting privacy mechanisms (see [57] for a survey). Within this body of work, there is a literature on privacy and *price discrimination* (e.g., [58–61]). We build on this work and extend to a setting that relaxes common-prior assumptions for buyers and sellers so that players must now devise strategies based on what they learn from repeated interactions.

In these repeated interactions, we observe the emergence of a *reputation-based privacy mechanism*. This reputation, learned by buyers based on previous interactions, takes the place of the prior that is used in the single-interaction game. There are numerous papers in economics on reputation focusing on sellers' reputations for the quality of the proffered good [62–64]. We focus on seller's reputation for enacting price discrimination and analyze how this arises in an online learning framework.

We also study the differences in behavior that arise from seller *commitment*, which has been studied in [65], [58], [61] and [66]. We show that even without commitment, similar behavior can arise through repeated interactions where reputation substitutes for the role of commitment.

Finally, we draw upon work on *online learning* and *repeated games*. There are a number of papers [67–70] on repeated interactions between a principal and an agent where the agent chooses actions based on evolving beliefs about the principal's actions. In our setting, we interpret the evolving beliefs as the reputation of the principal. Our setting differs in two ways. The first is that the principal's actions are not revealed at the end of the round. Instead partial information about the action, depending on the agent's response, is revealed. The second is that our results hold for weaker conditions on the agent's beliefs compared to previous work.

4.2 A Price-Discrimination Game

We formulate price discrimination as a sequential, incomplete-information game between n buyers and a seller.

Definition 4 (PD game)**.** *The price-discrimination game with parameters $n, \alpha, \mu, \overline{\theta}, \underline{\theta}, c_B, c_S$, denoted the* $(n, \alpha, \mu, \overline{\theta}, \underline{\theta}, c_B, c_S)$-PD game*, has the following extensive-form representation.*

1. ***Nature's move.*** *The game begins with Nature assigning types to each participant according to random draws. For $i \in [n]$, the type for buyer i is $\theta_i \in \{\underline{\theta}, \overline{\theta}\}$, representing their valuation of the item being sold, with $\underline{\theta} < \overline{\theta}$. A buyer is type $\overline{\theta}$ with probability μ and type $\underline{\theta}$ with probability $1 - \mu$. The seller's type χ is either* signal aware *($\chi = 1$) or* signal blind *($\chi = 0$). The seller is signal aware with probability α and signal blind with probability $1 - \alpha$.*

2. ***Signaling stage.*** *Based on their assigned type θ_i, each buyer signals $s_i \in \{\underline{s}, \overline{s}\}$. Signaling one's true type ($\underline{s}$ for type $\underline{\theta}$ and $\overline{s}$ for type $\overline{\theta}$) incurs no cost, whereas signaling a mismatched type, referred to as "evasion," imposes a cost c_B on the buyer and a cost c_S on the seller.*[1]

3. ***Pricing decision.*** *The seller chooses a price p_i to set for buyer i. The information the seller can use to set the prices depends on the type of seller. A signal-aware seller can set prices depending on the signals sent by the buyers, that is, they can set one price for all buyers that signaled $\underline{s}$ and a different price for all buyers that signaled $\overline{s}$. A signal-blind seller must set the same price for all buyers since they have no information to distinguish buyers.*

4. ***Purchase decisions.*** *Each buyer, based on the price p_i set for them and their valuation θ_i, makes a choice $b_i \in \{0, 1\}$, to purchase the item ($b_i = 1$) or not ($b_i = 0$).*

5. ***Utilities.*** *All players receive their respective utilities. Each buyer's positive utility is zero if they do not buy the item and the difference between their valuation and price otherwise. If they took evasive action in the signaling stage, their negative utility is equal to their cost of evasion c_B. That is, buyer i's utility is*

$$u_B(\theta_i, s_i, p_i, b_i) = (\theta_i - p_i)b_i - c_B e(\theta_i, s_i) \tag{4.1}$$

where $e(\theta_i, s_i) = \mathbb{1}\{(\theta_i = \underline{\theta} \wedge s_i = \overline{s}) \vee (\theta_i = \overline{\theta} \wedge s_i = \underline{s})\}$ indicates evasion or not. The seller's overall utility is the sum of utilities $u_S(\theta_i, s_i, p_i, b_i)$ from their interactions with each buyer. The positive utility due to buyer i is the revenue p_i if buyer i buys and zero otherwise. If the buyer took evasive action in the signaling stage, the seller incurs negative utility c_S. That is, the seller's utility is

$$u_S\left((\theta_i, s_i, p_i, b_i)_{i=1}^n\right) = \sum_{i=1}^n u_S(\theta_i, s_i, p_i, b_i) = \sum_{i=1}^n p_i b_i - c_S e(\theta_i, s_i). \tag{4.2}$$

[1] We can more generally allow for each type of buyer impose a different evasion cost (e.g., if a $\overline{\theta}$-buyer evades, the costs are $\bar{c}_B, \bar{c}_S \in \mathbb{R}$, and if a $\underline{\theta}$-buyer evades, the costs are $\underline{c}_B, \underline{c}_S \in \mathbb{R}$. However, as we later show, the only costs that are relevant are the evasion costs associated with the $\overline{\theta}$-seller, because the $\underline{\theta}$ seller will never choose to evade, so we can think of $c_B = \bar{c}_B$ and $c_S = \bar{c}_S$.

Mixed strategies. For simplicity of presentation, our game definition is stated in terms of pure strategies (i.e., players take deterministic actions). However, we can more generally allow players to employ mixed strategies. A *mixed strategy* for a player is a distribution over allowed actions conditioned on the information available when taking the action: buyer i's mixed signaling strategy induces a conditional distribution over signals $\pi_i^s(\cdot|\theta_i) \in \Delta(\{\underline{s}, \overline{s}\})$; the seller's mixed pricing strategy induces conditional distributions $\pi^p(\cdot|\underline{s}, \chi)$, $\pi^p(\cdot|\overline{s}, \chi)$ over positive reals with the constraint $\pi^p(\cdot|s = \underline{s}, \chi = 0) = \pi^p(\cdot|s = \overline{s}, \chi = 0)$; finally, each buyer i's mixed buying strategy induces conditional distribution $\pi_i^b(\cdot|\theta_i, p_i) \in \Delta(\{0, 1\})$.

Let $\pi = (\pi^s, \pi^p, \pi^b)$ denote a mixed strategy profile. π, along with the probability of player types described in Step 1 of Definition 4 (which we will denote $p(\chi)$ and $p(\theta_i)$) induce a distribution over action profiles with the probability of an action profile $(\chi, (\theta_i, s_i, p_i, b_i)_{i=1}^n)$ given by

$$\mathbb{P}\left(\chi, (\theta_i, s_i, p_i, b_i)_{i=1}^n\right) = p(\chi) \prod_{i=1}^n p(\theta_i) \pi_i^s(s_i|\theta_i) \pi^p(p_i|\theta_i, \chi) \pi_i^b(b_i|\theta_i, p_i). \tag{4.3}$$

Given a mixed strategy profile π, we will denote the expected utility for the seller and buyer i by

$$U_S(\pi) = \mathbb{E}\left[u_S\left((\theta_i, s_i, p_i, b_i)_{i=1}^n\right)\right] \qquad \text{and} \qquad U_B^i(\pi) = \mathbb{E}\left[u_B\left(\theta_i, s_i, p_i, b_i\right)\right],$$

where the expectation is over the joint distribution in (4.3).

Solution concept. We study the *perfect Bayes Nash equilibrium (PBNE)*. Mixed strategies of players constitute a PBNE if the following conditions hold: (1) sequential rationality, meaning that each player's strategy constitutes a best response to their beliefs about the other players' types and strategies, given the history of the game up to the point of choosing the action and (2) consistency of beliefs, meaning that players' beliefs about other players' types are updated following Bayes' rule.

The following theorem characterizes the PBNE of the price-discrimination game described in Definition 4.

Theorem 6. *An $(n, \alpha, \mu, \overline{\theta}, \underline{\theta}, c_B, c_S)$-PD game has the following unique perfect Bayes Nash equilibrium. Define $\Delta\theta = \overline{\theta} - \underline{\theta}$.*

(a) Buyers with type $\theta_i = \underline{\theta}$ will signal $s_i = \underline{s}$.

(b) Buyers with type $\theta_i = \overline{\theta}$ will signal

$$s_i = \begin{cases} \underline{s} \text{ w.p. } q^* & \text{if } \alpha > c_B/\Delta\theta \\ \overline{s} & \text{otherwise.} \end{cases} \qquad \text{where } q^* = \min\left\{1, \frac{(1-\mu)\theta}{\mu\Delta\theta}\right\}$$

(c) The signal-aware seller sets price

$$p^*_{\textit{signal aware}}(s) = \begin{cases} \underline{\theta} & \textit{if signal } s = \underline{s} \textit{ is observed} \\ \overline{\theta} & \textit{if signal } s = \overline{s} \textit{ is observed.} \end{cases}$$

(d) The signal-blind seller sets price

$$p^*_{\textit{signal blind}} = \begin{cases} \underline{\theta} & \textit{if } \underline{\theta} \geq \mu\overline{\theta} \\ \overline{\theta} & \textit{otherwise.} \end{cases}$$

(e) Buyer i buys the good if and only if their price p_i is at most their value, so

$$b_i = \mathbb{1}\{\theta_i \leq p_i\}.$$

The proof is given in Appendix 4.7.

Remark 2 (Buyer-induced privacy)**.** *The $\overline{\theta}$-buyers' equilibrium response can be interpreted as a privacy-protecting mechanism. This type of buyer is vulnerable to price discrimination, so rather than always signaling their true type, they may choose to randomize their signal. More specifically, if the cost of evasion is very high, the $\overline{\theta}$-buyer will tell the truth, but if the evasion cost is low enough, the $\overline{\theta}$-buyer can receive a reduction in price that is higher than their evasion cost. In the latter case, the $\overline{\theta}$-buyer must then choose the maximum evasion probability q^* such that it is still in the seller's best interest to take the buyer's signal at face value. We call this randomization "buyer-induced privacy."*

Theorem 6 tells us that strategic behavior can only happen if $c_B < \Delta\theta$ (otherwise, we can never have $\alpha > {}^{c_B}/_{\Delta\theta}$, so buyers will always signal truthfully). For the rest of the paper, we will focus on this setting.

Assumption 1. *In all following results, we assume $c_B < \Delta\theta$.*

A natural next question is how each player's utility is affected by the game parameters. In particular, we focus on the effect of α, due to its connection to privacy. In Figure 4.1, we visualize the utilities of the seller and $\overline{\theta}$-buyers as α varies from 0 to 1. Observe that the seller's utility increases for α less than some threshold value α^*, whose exact value we give in the corollary below. This corresponds to the set of PD-games where the buyer's equilibrium response is truthful. Beyond α^*, the $\overline{\theta}$-buyers' equilibrium response changes to being strategic and the seller's utility drops. We formalize the ordering of utilities in the following corollary.

Corollary 2. *(Order of utilities) Fix $n, \mu, \overline{\theta}, \underline{\theta}, c_B, c_S$ and let $u_S(\alpha), u_B(\alpha)$ denote the seller's and $\overline{\theta}$-buyers' equilibrium utilities of the $(n, \alpha, \mu, \overline{\theta}, \underline{\theta}, c_B, c_S)$-PD game. $u_S(\cdot)$ is maximized at $\alpha^* = c_B/\Delta\theta$, and the equilibrium utilities for the settings where the seller is always signal blind ($\alpha = 0$), is always signal aware ($\alpha = 1$), and is signal aware with probability α^* ($\alpha = \alpha^*$) have the following orderings:*

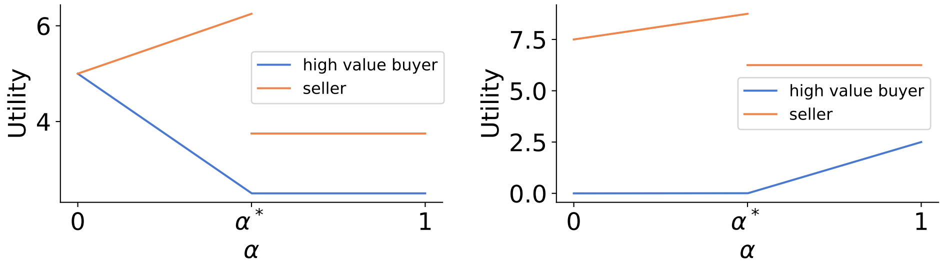


Figure 4.1: Plots of the $\overline{\theta}$-buyer and seller utilities as a function of α in the $\underline{\theta} \geq \mu\overline{\theta}$ setting (left) and the $\underline{\theta} < \mu\overline{\theta}$ setting (right).

(a) When $\underline{\theta} \geq \mu\overline{\theta}$,

$$u_S(\alpha^*) > u_S(0) > u_S(1) \qquad \textit{and} \qquad u_B(0) > u_B(1) = u_B(\alpha^*).$$

(b) When $\underline{\theta} < \mu\overline{\theta}$,

$$u_S(\alpha^*) > u_S(0) > u_S(1) \qquad \textit{and} \qquad u_B(1) > u_B(0) = u_B(\alpha^*).$$

$\underline{\theta}$-buyers always receive a utility of zero, regardless of the value of α.

The proof is given in Appendix 4.7.

Price discrimination with seller commitment

A key takeaway from Corollary 2 is that the seller's utilities are dependent on the value of α, and if the seller could choose a value of α, they would want to choose $\alpha = \alpha^*$ to maximize their utility. Suppose we are now in a setting where the seller is able to choose and publicly commit to an α. As a motivating example, suppose that the seller must go through a data broker to access signals, and the data broker publishes trusted summaries of what fraction of buyers the seller requests data on. In such a setting, where α is chosen by the seller instead of treated as given, we arrive at the following equilibrium.

Corollary 3. *(Equilibrium of price-discrimination game with commitment) When the seller has commitment power (i.e., is able to credibly communicate to sellers that they will not price discriminate with some probability), the perfect Bayes Nash equilibrium of the PD-game consists of the following strategies:*

(a) The seller commits to not price-discriminating (by playing $p^*_{signal\ blind}$ from Theorem 6) with probability $1 - \alpha^*$, where $\alpha^* = c_B/\Delta\theta$.

(b) All buyers always signal truthfully.

The buyers' buying decisions are the same as in Theorem 6.

Proof. (a) follows directly from Corollary 2, which tells us that the seller's utility is maximized at α^*, and (b) comes from applying Theorem 6 with $\alpha = \alpha^*$. □

Remark 3. *Commitment ability allows the seller to achieve a higher utility by providing seller-induced privacy. This seller-induced privacy obviates the need for buyers to take evasive action to create buyer-induced privacy, which benefits the seller. We use $\mathbb{U}_1^*$ to refer to the seller's maximum achievable equilibrium utility in the single interaction price discrimination game with commitment. This utility is achieved when the seller plays the strategy given in Corollary 3.*

4.3 Repeated Interactions

In the previous section, we saw the emergence of seller-induced privacy when the seller has commitment ability. If possible, the seller would commit to providing seller-induced privacy (by ignoring signals with probability $1 - \alpha^*$, as in Corollary 3), thereby limiting the extent of price discrimination performed by the seller. However, these results hinge on the buyer believing that the α stated by the seller truly corresponds to the probability of price discrimination. Without this credible commitment from the seller, the story becomes more complicated.

In this section, we study whether seller-induced privacy can still arise in the absence of such commitment ability, through the development of a reputation based on the seller's historical pricing. We ask the question of how the extent of privacy and resulting utilities differ under reputation-based privacy versus commitment-based privacy. We model the seller as making pricing decisions using an online learning algorithm and show how different models such as *no-regret* and *no-policy-regret* lead to different answers to this question.

In the repeated interaction setting, we also relax the assumptions that the distribution μ over agent types and the probability α that the seller looks at the agent's signal are publicly known. Rather than playing the single-interaction equilibrium strategies, which require full knowledge of game parameters, the players now have to learn strategies online based on past interactions.

Setup

We consider repeated interactions between a seller and buyers where a new batch of buyers is drawn at each round. We call this as the *repeated PD protocol*. Each round is similar to the one-shot PD-game from Definition 4 but with the following differences: (1) There is one fixed seller throughout all rounds. (2) When players choose actions, they not only have access to information from the current round (as was the case in the one-shot PD game) but also some information from previous rounds. Specifically, at round t, the seller has access to

$((s_i^\tau, p_i^\tau)_{i=1}^n)_{\tau=1}^{t-1}$, the signals they observed and the prices they set in previous rounds, and each buyer i has access to $(((\theta_i^\tau, s_i^\tau, p_i^\tau)_{i=1}^n)_{\tau=1}^{t-1})$, the buyer types, signals, and prices of all buyers from previous rounds. This modeling of the buyers' access is appropriate in settings where buyer information is pooled either through crowd-sourcing or by an auditing entity and made available to buyers. (3) The parameter μ (the probability of a type-$\overline{\theta}$ buyer) is not known to the seller. (4) The probability that the seller will price discriminate is not known to buyers, as was assumed in the one-shot PD game; rather, buyers must estimate this probability based on past rounds. We write out the repeated interaction protocol in detail in Appendix 4.6.

Model of the buyers

Since each buyer participates in only one round of the repeated PD protocol, the equilibrium response is still appropriate to model the buyer's response. However, in the repeated interaction setting, we no longer assume the buyers hold a static, prior belief about the probability of a signal-aware seller. Instead, buyers have evolving beliefs based on the seller's interactions with past buyers.

Some specific buyer strategies we will refer to are π^s_{truthful}, which corresponds to always signaling truthfully, and $\pi^s_{\text{strategic}}$, which corresponds to signaling $\underline{s}$ with probability q^* (as defined in Theorem 6) and signaling $\overline{s}$ with probability $1 - q^*$. We consider the following model of buyer behavior.

Definition 5 (Consistent belief based equilibrium responding (CBER) buyers)**.** *Consistent belief based equilibrium responding buyers (or CBER-buyers) form a sequence of beliefs $(\hat{\alpha}_t)_{t=1}^T$ satisfying a consistency property defined below and at round t, choose the corresponding equilibrium strategy (from Theorem 6) of the PD-game with $\alpha = \hat{\alpha}_t$. That is, $\underline{\theta}$-buyers always signal truthfully, and $\overline{\theta}$-buyers signal truthfully (play $\pi^s_{truthful}$) if $\hat{\alpha}_t \leq \alpha^*$ and signal the opposite type with probability q^* otherwise (play $\pi^s_{strategic}$).*

We now explain the consistency property. Given a sequence of seller mixed strategies action profiles that induce the sequences of distributions $(\pi_t^p(\cdot|s = \overline{s}))_{t=1}^T$ and $(\pi_t^p(\cdot|s = \underline{s}))_{t=1}^T$ indicating price distributions at each round for signals $\underline{s}, \overline{s}$ respectively, define α_t to be

$$\alpha_t = \mathbb{P}_{\overline{P}\sim\pi_t^p(\cdot|s=\overline{s}),\underline{P}\sim\pi_t^p(\cdot|s=\underline{s})}\left[\overline{P} \neq \underline{P}\right].$$

That is, α_t denotes the probability of a different price for $\overline{s}$ compared to $\underline{s}$ at round t. The probability here is over the randomness due to the seller's mixed strategy at round t. α_t is a measure of extent of price discrimination by the seller at round t.

Definition 6 (Consistent sequence)**.** *Let $\bar{\alpha}_T = (1/T)\sum_{t=1}^T \alpha_t$. We say a sequence of estimators $(\hat{\alpha}_t)_{t=1}^T$ is* consistent *if $\lim_{T\to\infty} |\mathbb{E}[\hat{\alpha}_T] - \bar{\alpha}_T| = 0$, where the expectation is taken over the randomness of the history $H_T = ((\theta_i^t, s_i^t, p_i^t)_{i=1}^n)_{t=1}^{T-1}$ used to construct $\hat{\alpha}_T$.*

A useful implication of consistency is that $\hat{\alpha}_T$ converges pointwise to $\bar{\alpha}_T$.

Lemma 26. *If $(\hat{\alpha}_t)_{t=1}^T$ is a consistent sequence of beliefs, then for any $\epsilon < 0$ and $\delta > 0$, there exists some positive integer N such that for all $T > N$, we have $\mathbb{P}\left[|\hat{\alpha}_T - \bar{\alpha}_T| \geq \epsilon\right] \leq \delta$.*

Proof. Due to consistency and the definition of limits, there exists N such that for all $T > N$, we have $|\mathbb{E}[\hat{\alpha}_T] - \bar{\alpha}_T| \leq \delta\epsilon$. Thus, for $T > N$, we can apply Markov's inequality to get $\mathbb{P}(|\hat{\alpha}_T - \bar{\alpha}_T| \geq \epsilon) \leq (|\mathbb{E}[\hat{\alpha}_T] - \bar{\alpha}_T|)/\epsilon \leq \delta\epsilon/\epsilon = \delta$. □

The following proposition and associated proof provide an algorithm to construct a consistent sequence of estimators $(\hat{\alpha}_t)_{t=1}^T$.

Proposition 6 (Existence of consistent sequence)**.** *Assume that buyers equilibrium-respond to $\hat{\alpha}_t$ at each round t. Then, for any sequence of seller actions, there exists a sequence of estimators $(\hat{\alpha}_t)_{t=1}^T$ that is consistent.*

Proof sketch: Since there are multiple buyers at each round, we can infer whether the seller is price discriminating or not by comparing the prices charged to a buyer who signals $\underline{s}$ and a buyer who signals $\overline{s}$. However, only some rounds are informative about price discrimination; in rounds where all buyers send the same signal, we are not able to determine if the seller had a price discriminatory pricing policy in place. The consistent estimator $\hat{\alpha}_t$ we consider is the fraction of past rounds where price discrimination is observed, normalized to account for the probability that a round is likely to be informative about price discrimination. We show that $E[\hat{\alpha}_t] = (1/t)\sum_{\tau=1}^{t-1}\alpha_\tau$, which implies that $\lim_{T\to\infty}|\mathbb{E}[\hat{\alpha}_T] - \bar{\alpha}_T| = \lim_{T\to\infty}\left|(1/T)\sum_{t=1}^{T-1}\alpha_t - (1/T)\sum_{t=1}^{T}\alpha_t\right| = \lim_{T\to\infty}\alpha_T/T = 0$. See Appendix 4.7 for the full proof.

Model of the seller

Since the seller does not a priori know the distribution over buyer types and is engaged in multiple rounds of the repeated interaction, modeling the seller's response by the one-shot equilibrium from Theorem 6 is not reasonable. Instead, we consider the seller as optimizing various common objectives of repeated interactions such as regret minimization and policy-regret minimization.

The seller's mixed strategy at a given round is a pair of probability distributions $\pi_t^p = (\pi_t^p(\cdot|\overline{s}), \pi_t^p(\cdot|\underline{s}))$. Let Π denote the set of possible mixed strategies. For rational sellers, we can focus on distributions supported only on $\{\underline{\theta}, \overline{\theta}\}$ without loss of generality. Prices supported on $\{\underline{\theta}, \overline{\theta}\}$ maximize seller revenue in each round. The seller's effect on future rounds is also not affected by limiting the support. This is because the parameters α_t that the buyers' consistent estimator estimates treats *any* difference in prices as indicating price discrimination, so all price differences are treated the same.

Some specific seller strategies we will refer to are π_{PD}^p and π_{noPD}^p. The former is the "always-price-discriminating strategy," with $\pi_{\text{PD}}^p(\overline{\theta}|\overline{s}) = \pi_{\text{PD}}^p(\underline{\theta}|\underline{s}) = 1$. The latter is the "never-price-discriminating strategy," with $\pi^p_{\text{noPD}}(\underline{\theta}|\underline{s}) = \pi^p_{\text{noPD}}(\underline{\theta}|\overline{s}) = 1$ if $\underline{\theta} \geq \mu\overline{\theta}$ and $\pi^p_{\text{noPD}}(\overline{\theta}|\underline{s}) = \pi^p_{\text{noPD}}(\overline{\theta}|\overline{s}) = 1$ otherwise.

Regret-minimizing seller

The first seller model we consider is a regret-minimizing seller.

Definition 7 (Seller's regret)**.** *Given a sequence of mixed strategy profiles* $\{\pi_t\} = \{(\pi^s_t, \pi^p_t, \pi^b_t)\}^T_{t=1}$*, the* seller's average regret *is*

$$R^S_T(\{\pi_t\}^T_{t=1}) = \frac{1}{T}\left[\max_{\pi^{p*}\in\Pi}\sum_{t=1}^T U_S(\pi^s_t, \pi^{p*}_t, \pi^b_t) - \sum_{t=1}^T U_S(\pi^s_t, \pi^p_t, \pi^b_t)\right].$$

Definition 8 (No-regret algorithm)**.** *Let $\mathcal{A}_B$ be an algorithm employed by the buyer in the repeated PD protocol. A seller algorithm $\mathcal{A}_S$ in the repeated PD protocol is a* no-regret algorithm *for the seller given $\mathcal{A}_B$ if the sequence of mixed strategies $(\pi_t)^T_{t=1}$ generated by the interaction between $\mathcal{A}_B$ and $\mathcal{A}_S$ has seller's average regret that is sublinear in the number of rounds. That is, $R^S_T((\pi_t)^T_{t=1}) \in o(1)$.*

We will denote by $(\pi_t)^T_{t=1}$ the sequence of random variables denoting the players' mixed strategies in each round. Our results analyze the asymptotic convergence of average seller utility. We say that the average seller utility *asymptotically converges* to some value v if $\lim_{T\to\infty}\mathbb{E}\left[(1/T)\sum_{t=1}^T U_S(\pi_t)\right] = v$. We write $U_S(\pi^p)$ and $U_S(\pi^s, \pi^p)$ when it is clear what the other arguments are.

If the seller employs a no-regret algorithm, then the seller could end up always price-discriminating i.e., no seller-induced privacy. This is stated below.

Proposition 7. *(Always price-discriminating is regret minimizing) Given CBER-buyers, the seller algorithm that always employs the price-discrimination strategy i.e., $\pi^p_t = \pi^p_{PD}$ for all timesteps t is a no-regret algorithm for the seller. The seller's average utility asymptotically converges to a value at most $u_S(1)$, where $u_S(1)$ is the seller's equilibrium utility in the single-interaction PD-game with $\alpha = 1$.*

Proof sketch. The strategy of CBER-buyers in each round is either π^s_{truthful} or $\pi^s_{\text{strategic}}$. For both these buyer responses, the seller's optimal strategy is to always price discriminate, as shown in the computation of the seller's equilibrium response in the proof of Theorem 6. In other words, the seller incurs zero regret in each round by always price-discriminating.

Next, we analyze the seller's average utility. Note that when $\pi^p_t = \pi^p_{\text{PD}}$, the probability of seeing different prices for different signals is $\alpha_t = 1$, so $\bar{\alpha}_t = 1$ for all t. By Lemma 26, $\hat{\alpha}_t$ becomes greater than α^* eventually (where α^* is as defined in Corollary 3), which causes $\overline{\theta}$-buyers to play $\pi^s_{\text{strategic}}$. In other words, eventually the seller and buyers will all be playing their equilibrium strategies for the PD-game with $\alpha = 1$, so their average utilities will converge to the corresponding equilibrium utilities. See Appendix 4.7 for the full proof. □

The next proposition tells us that regret minimization necessarily causes the seller to achieve a worse expected average utility that the optimal utility they can achieve in the single interaction setting.

Proposition 8 (Regret minimization is inherently at odds with achieving $\mathbb{U}_1^*$)**.** *Given CBER-buyers, for any no-regret seller algorithm, the seller's average utility asymptotically converges to strictly less than* $\mathbb{U}_1^*$.

Proof sketch. Define $\mathcal{T} = \{t \in [T] : \hat{\alpha}_t \leq \alpha^*\}$ to be the set of rounds where $\overline{\theta}$-buyers' signaling strategy is π^s_{truthful}. In all other rounds, their signaling strategy is $\pi^s_{\text{strategic}}$. Define $\beta = (1/T)\sum_{t\in\mathcal{T}}\alpha_t$ to be a measure of simultaneous truthfulness from buyers and price-discrimination by the seller. Our proof involves the following parts. We outline the parts and state them as lemmas here and prove them in Appendix 4.7

1. Obtaining $\mathbb{U}_1^*$ requires the buyers to be truthful strictly more than α^* fraction of rounds.

   Lemma 27. $\lim_{T\to\infty} |\mathcal{T}|/T \leq \alpha^*$ *implies that* $\lim_{T\to\infty}\left(\sum_{t=1}^T U_S(\pi_t)\right)/T < \mathbb{U}_1^*$.

2. The no regret property requires that the seller price discriminates in most rounds where buyers are truthful. So β is close to $|\mathcal{T}|/T$.

   Lemma 28. $\lim_{T\to\infty} |\mathcal{T}|/T \leq \lim_{T\to\infty}\sum_{t\in\mathcal{T}}\alpha_t/T$.

3. There is a limit on simultaneous price-discrimination and truthful signaling due to the buyers' consistent beliefs. That is, β converges to at most α^*.

   Lemma 29. $\lim_{T\to\infty}\sum_{t\in\mathcal{T}}\alpha_t/T \leq \alpha^*$.

From Lemmas 28, 29, $\lim_{T\to\infty} |\mathcal{T}|/T \leq \alpha^*$. Lemma 27 shows that this means average seller utility is strictly less than $\mathbb{U}_1^*$.

□

Policy-regret-minimizing seller

As we have seen, regret minimization does not guarantee that the seller achieves higher than price-discrimination utility. On the other hand, if we model the seller as minimizing policy regret [71], the seller *necessarily* achieves utility that is higher than the utility achieved by the naive strategy of always price discriminating.

Definition 9 (Seller's policy regret)**.** *Consider a buyer algorithm $\mathcal{A}_B$ and a seller algorithm $\mathcal{A}_S$. Let $(\pi_t(\mathcal{A}_B, \mathcal{A}_S))_{t=1}^T$ be the sequence of mixed strategies generated by the interaction between $\mathcal{A}_B$ and $\mathcal{A}_S$. Given a sequence of mixed strategies $(\pi_t)_{t=1}^T$, the* seller's average policy regret *of $(\pi_t)_{t=1}^T$ relative to a buyer algorithm $\mathcal{A}_B$ and a baseline class $\mathbb{A}_S$ of seller algorithms is*

$$PR_T^S\left((\pi_t)_{t=1}^T; \mathcal{A}_B, \mathbb{A}_S\right) = \max_{\mathcal{A}_S \in \mathbb{A}_S} \frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t(\mathcal{A}_B, \mathcal{A}_S)) - \frac{1}{T}\sum_{t=1} U_S(\pi_t)$$

Definition 10 (No-policy-regret algorithm)**.** *Let $\mathcal{A}_B$ be an algorithm employed by the buyer in the repeated PD protocol. An algorithm $\mathcal{A}_S$ is a* no-policy-regret algorithm *for the seller given $\mathcal{A}_B$ and relative to a class of seller algorithms $\mathbb{A}_S$ if the sequence of mixed strategies $(\pi_t(\mathcal{A}_B, \mathcal{A}_S))_{t=1}^T$ generated by the interaction between $\mathcal{A}_B$ and $\mathcal{A}_S$ satisfies $PR_T^S((\pi_t(\mathcal{A}_B, \mathcal{A}_S); \mathcal{A}_B, \mathbb{A}_S)_{t=1}^T) \in o(1)$.*

Consider a baseline class $\mathbb{A}_S^{MS}$ consisting of seller algorithms that employ the same mixed strategy in each round, that is, $\pi_t^p(\cdot|\overline{s})$ is the same distribution for all t and similarly for $\pi_t^p(\cdot|\underline{s})$.

Proposition 9 (Policy-regret-minimizing seller achieves $\mathbb{U}_1^*$)**.** *Given CBER-buyers, if the seller achieves sub-linear policy regret relative to $\mathbb{A}_S^{MS}$, then the seller's average utility asymptotically converges to at least $\mathbb{U}_1^*$.*

Proof sketch. Under the conditions of this proposition, the seller's utility must, by definition of policy regret, approach a utility at least as high (or better) than the utility of any strategy in $\mathbb{A}_S^{MS}$ as $T \to \infty$. Recall that $\mathbb{U}_1^*$ is the seller utility achieved in the PD game when $\alpha = \alpha^*$. Consider the PD game that results in a seller utility of at least $\mathbb{U}_1^* - \epsilon$, which is achieved by the seller price-discriminating with probability $\tilde{\alpha} < \alpha^*$. Then the repeated-interaction strategy of always price-discriminating with probability $\tilde{\alpha}$ has an average expected utility of at least $\mathbb{U}_1^* - \epsilon$ (this must be true due to the consistency of buyer beliefs; see the full proof in Appendix 4.7 for details). Taking ϵ to 0 gives the desired result. □

Combining the previous result with the following result tells us that a no-policy regret seller's algorithm will cause the seller's average utility to asymptotically converge to *exactly* $\mathbb{U}_1^*$. In fact, this result tells us the stronger result that there does not exist *any* seller algorithm that can achieve utility higher than $\mathbb{U}_1^*$.

Proposition 10. *Given CBER-buyers, for any seller algorithm, the seller's average utility asymptotically converges to at most $\mathbb{U}_1^*$.*

Proof sketch. This proof is similar to the argument of the proof of Proposition 8 and the full proof is in Appendix 4.7. The key ideas again are that for high seller utility, there must be sufficiently many rounds where simultaneously, the seller price discriminates and the buyer reports truthfully. Since the buyers' belief estimators are consistent, this cannot be the case. The difference between the average seller utility and $\mathbb{U}_1^*$ is a constant times the following quantity: $\frac{1}{T}\sum_{t\in\mathcal{T}}(\pi_t^p(\overline{\theta}|\overline{s}) - \pi_t^p(\underline{\theta}|\underline{s})) - \alpha^*$, where $\mathcal{T}$ is the set of rounds where the buyer signals truthfully. Lemma 30, 29 (from the proof of Proposition 8) show that the consistency property implies that this difference converges to most zero. □

4.4 Experiments

In this section, we simulate Algorithm 6 with $\mu = 0.5$, $\underline{\theta} = 5$, $\overline{\theta} = 15$, $c_B = c_S = 5$, and $n = 10$ and empirically verify our theoretical claims from Section 4.3. We report the convergence of buyer and seller utilities, seller actions, and buyer estimators. The seller and buyer algorithms we consider are described below. Code for all experiments is available at this github repo.

Algorithms

Seller

1. **Signal-blind seller.** The seller plays the regret-minimizing Exp3 algorithm (specifically Exp3-IX in Chapter 12 of [72]). At round t the seller sets a price $p_t \in \{\underline{\theta}, \overline{\theta}\}$ according to the algorithm's current sampling distribution, charges p_t to all buyers and updates the sampling distribution based on the resulting average utility from the buyers' purchase decisions.

2. **Signal-aware seller.** The seller plays a contextual version of Exp3, which we call CExp3, in which the algorithm maintains two sampling distributions over prices $\{\underline{\theta}, \overline{\theta}\}$, conditioned on the received signal, $\underline{s}$ or $\overline{s}$. At each round, the seller samples once from each distribution and charges one price $\underline{p}_t$ to all buyers who signal $\underline{s}$ and $\overline{p}_t$ to all buyers who signal $\overline{s}$. Depending on the sampling distributions, $\underline{p}_t$ and $\overline{p}_t$ may or may not be equal.

3. **Stackelberg equilibrium seller.** The seller commits to an $\alpha^* = {}^{c_B}/_{\Delta\theta}$ level of price-discrimination, i.e., they play the $(\alpha = 1)$-PD equilibrium strategy (Theorem 6) with probability α^* and the $(\alpha = 0)$-PD equilibrium with probability $1 - \alpha^*$.

CBER-Buyer. Using a sequence of consistent estimators $\{\hat{\alpha}_\tau\}_{\tau=1}^{t-1}$ (Def. 6) to estimate the seller's probability of price-discrimination at each round, each buyer plays the $(\alpha = \hat{\alpha}_\tau)$-PD equilibrium strategy. For our simulations, buyers use the estimator described in (4.7) to estimate the seller's probability of price discrimination at each round. All buyers in a single round use the same estimator.

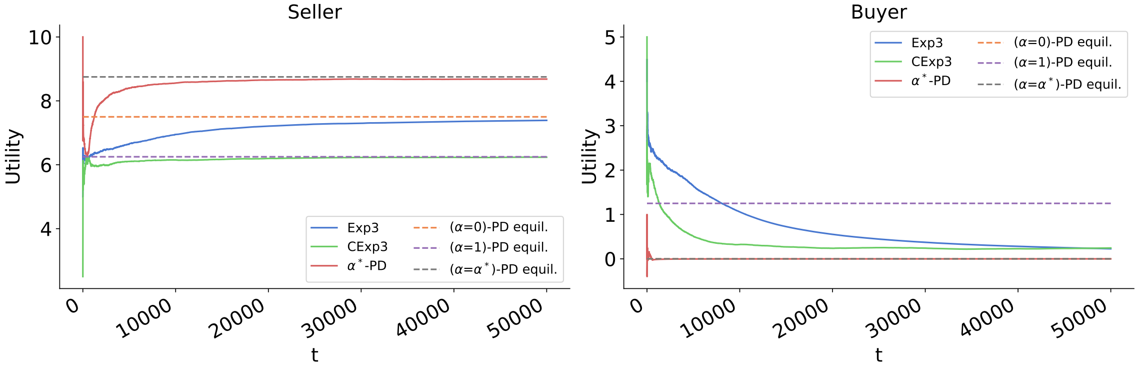


Figure 4.2: Convergence of seller and buyer utilities for various algorithms. $\underline{\theta} < \mu\bar{\theta}$ with our experiment parameters, so the buyer's $(\alpha = 0)$-PD and $(\alpha = \alpha^*)$-PD utilities are the same (see Corollary 2).

Discussion

Convergence of Utilities. Figure 4.2 shows convergence of seller and buyer utilities for each of the seller's algorithms played against a CBER-buyer. As expected, when a seller plays Exp3 (which ignores signals) against a CBER-buyer, the players' utilities converge to the $(\alpha = 0)$-PD equilibrium utility (Theorem 6). When the seller plays CExp3 (which observes signals) against a CBER-buyer, the seller's utility converges to the $(\alpha = 1)$-PD equilibrium utility. Given our experiment parameters, multiple different distributions $\pi_t^p(\cdot|s = \underline{s})$ reward the seller equivalently, while some are more favorable for the buyer than others. Therefore, while the seller's utility will always converge to $(\alpha = 1)$-PD, the buyer's utility may converge to something less than $(\alpha = 1)$-PD. Finally, when the seller plays the Stackelberg equilibrium against a CBER-buyer, the players' utilities converge to the $(\alpha = \alpha^*)$-PD equilibrium utility.

Consistency of $\hat{\alpha}$. Figure 4.3 illustrates the consistency of the buyer's estimator ((4.7)). Our simulations show that the buyer's estimate $\hat{\alpha}_t$ of the seller's probability of price discrimination converges to 0 against a seller playing Exp3, to 0.5 against a seller playing α^*-PD (where $\alpha^* = {}^{c_B}/_{\Delta\theta} = 0.5$ given our simulation parameters), and to higher-than-0.5 against a seller playing CExp3. Importantly, $\hat{\alpha}_t$ aligns with the seller's true average probability of price-discrimination, $\bar{\alpha}_t$, giving empirical evidence for Lemma 26.

Convergence of Seller Actions. In Figure 4.4, we track the cumulative proportion of the seller's price-discriminatory vs. non-price-discriminatory actions. Specifically, we track four seller actions: 1) charging a high price regardless of signal, 2) charging a low price regardless of signal, 3) charging a high price for a low signal and low price for a low signal (PD), and 4) charging a low price for a high signal and a high price for a low signal (reversePD). Given our parameter values for these simulations (i.e. $\underline{\theta} < \mu\bar{\theta}$ and $\alpha^* = 0.5$),

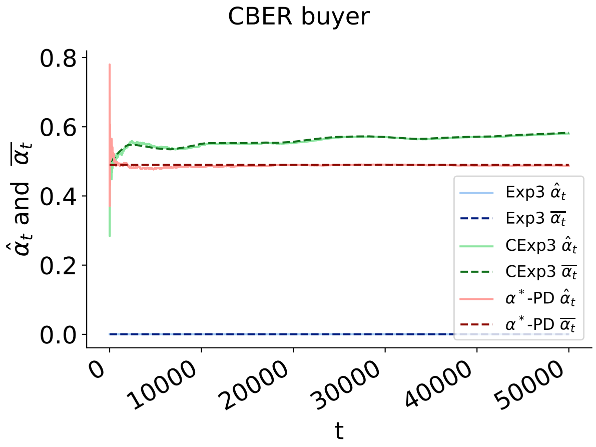


Figure 4.3: $\hat{\alpha}_t$ and $\bar{\bar{\alpha}}_t$ over time when seller is playing Exp3, CExp3, or α^*-PD. In all cases, $\hat{\alpha}$ is a consistent estimator of the seller's true probability of price discrimination.

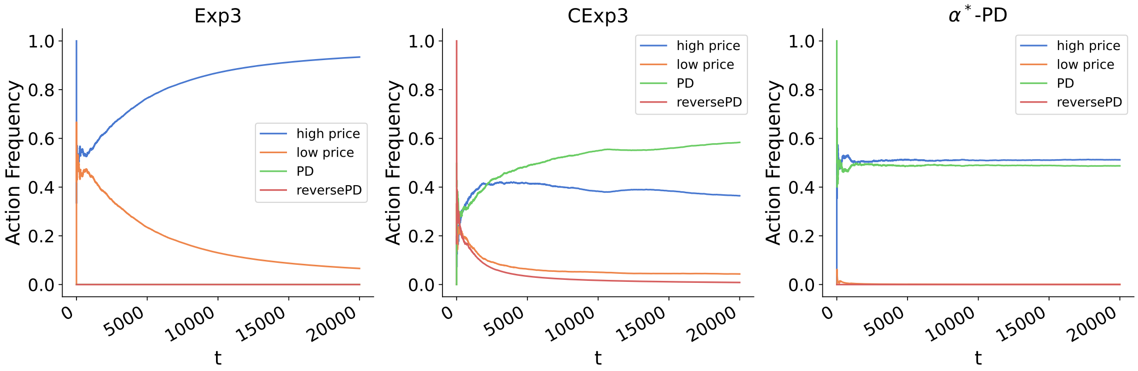


Figure 4.4: Relative frequency of actions for the seller playing Exp3, CExp3 and α^*-PD. The number of PD and reversePD actions for the Exp3 seller are both 0, as is expected.

in equilibrium we would expect that, for each batch of n buyers at a single round: 1) a signal-blind seller sets a high price for all n buyers, 2) a signal-aware seller sets a high price for high-signal buyers and a low price for low-signal buyers, and 3) a α^*-PD seller sets a high price for all high-signal buyers and low price for all low-signal buyers with probability 0.5 and sets a high price for all n buyers with probability 0.5. Figure 4.4 gives empirical evidence for this intuition.

Biased $\hat{\alpha}$. In realistic settings, the buyer may not have a consistent estimate of price discrimination and instead only have access to a biased $\hat{\alpha}$. Figure 4.5 examines whether a seller can benefit from non-consistency in the buyer's estimate. The y-axis of the figure tracks

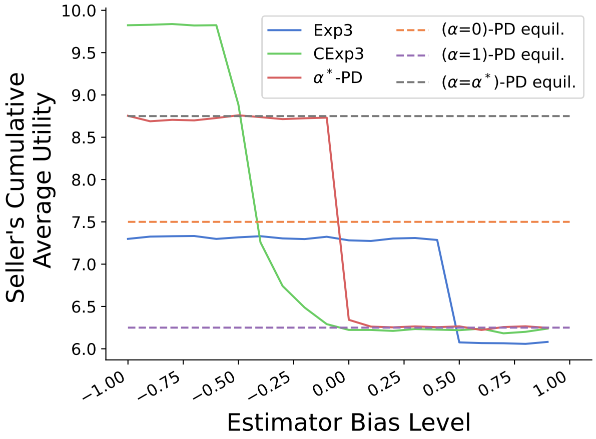


Figure 4.5: Cumulative average utility of the seller playing against CBER-buyers using biased $\hat{\alpha}$'s.

the seller's cumulative average utility after $20,000$ rounds of interaction with CBER-buyers. We partition the interval $[-1, 1]$ into twenty segments γ_i of width 0.1, and the buyers use estimator $\hat{\alpha}_t + \epsilon_t$, where $\epsilon_t \sim \text{Unif}(\gamma_i)$. The plot then tracks the seller's cumulative average utility after $20,000$ rounds of interaction with buyers for each biase interval γ_i. If $\hat{\alpha}_t + \epsilon_t$ is less than 0 or greater than 1, we clip it at those values respectively. In all cases, the seller is hurt by a $\overline{\theta}$-buyer who overestimates the probability of price discrimination (high values of ϵ_t) and is thus more likely to evade, costing the seller the evasion cost. Against a buyer who underestimates the probability of price discrimination (low values of ϵ_t), neither the Exp3 nor α^*-PD seller gains utility, since the equilibrium behavior of the buyer with consistent $\hat{\alpha}_t$ aligns with the no-price-discrimination equilibrium (see Figure 4.2). By contrast, the CExp3 seller benefits from a buyer who underestimates the probability of price discrimination, since the seller benefits from discriminatory pricing without incurring the evasion cost. Against a CBER-buyer with consistent estimates, this advantage is impossible at equilibrium.

4.5 Conclusion

Since the type and level of privacy desired generally depends on the utilities of stakeholders and forms of interaction among them, we propose a game theoretic framework for privacy in this paper. We analyzed the perfect Bayes Nash equilibrium in a single-interaction setting as well as no-regret and no-policy-regret dynamics emerging over repeated interactions. In both these settings, we show how the different components of the game—utilities, actions and information sets (information available to players when choosing actions) impact the privacy levels that emerge.

Our results shed light on the impacts of different privacy-related interventions—we showed that enabling a seller to credibly commit to privacy (e.g., through privacy legislation like the GDPR) or revealing the seller's past behavior (e.g., through privacy auditing) can surprisingly improve their utility. Thus, we believe our framework can be used to help analyze and craft privacy policies.

4.6 Repeated PD Protocol

Algorithm 6 makes clear what information is available to a given player at each point in the repeated interaction setting.

Algorithm 6 Repeated PD protocol

```
Parameters: $(\mu, \underline{\theta}, \overline{\theta}, c_B, c_S)$
$H_1^B, H_1^S := \varnothing$
for $t = 1$ to $T$ do
    for each buyer $i = 1$ to $N$ do
        Nature draws $\theta_i^t$ with $\theta_i^t = \overline{\theta}$ with probability $\mu$ and $\theta_i^t = \underline{\theta}$ otherwise
        Buyer $i$ chooses mixed strategy $\pi_S^i$ over signal $s_i^t \in \{\underline{s}, \overline{s}\}$ based on $\{\theta_i^t\} \cup H_t^B$
        Seller chooses mixed strategy $\pi^p(\cdot \mid s = s_i^t)$ over price $p_i^t$.
        Buyer $i$ decides to buy, denoted by indicator $b_i^t$, based on $\{\theta_i^t, p_i^t\} \cup H_t^B$
    Buyer $i$ receives utility $u_B(\theta_i^t, s_i^t, p_i^t, b_i^t)$ and seller receives utility $u_S\left((\theta_i^t, s_i^t, p_i^t, b_i^t)_{i=1}^n\right)$,
    as defined in (4.1) and (4.2).
    $H_{t+1}^B = H_t^B \cup \{(\theta_i^t, s_i^t, p_i^t)_{i=1}^n\}$
    $H_{t+1}^S = H_t^S \cup \{(s_i^t, p_i^t)_{i=1}^n\}$
```

4.7 Proofs of Theoretical Results

Proof of Theorem 6

Proof. Part (a) comes from the fact that $\underline{\theta}$ buyers have no reason to pretend to have a higher valuation for the good than they actually do. Part (e) comes from the fact that buyers are utility maximizing.

Part (c) comes from the following reasoning: since signal blind sellers cannot see the buyers' signals, they must choose one price to set for all buyers. The seller wants to maximize their revenue, so they would ideally want to set the highest price that the buyer is willing to pay ($\overline{\theta}$ for $\overline{\theta}$-buyers and $\underline{\theta}$ for $\underline{\theta}$-buyers). However, the seller does not know the type of the buyer; all they know is the probability μ that the buyer is $\overline{\theta}$. The seller has to make a decision between charging $\overline{\theta}$ or $\underline{\theta}$. If the seller charges $\underline{\theta}$, both $\underline{\theta}$ and $\overline{\theta}$ agents would be willing to buy, so the expected revenue is $\underline{\theta}$. If the seller charges the higher price $\overline{\theta}$, only $\overline{\theta}$ agents would be willing to buy, so the expected revenue is $\mu\overline{\theta}$, which corresponds to

$$p^*_{\text{signal blind}} = \begin{cases} \underline{\theta} & \text{if } \underline{\theta} \leq \mu\overline{\theta} \\ \overline{\theta} & \text{if } \underline{\theta} > \mu\overline{\theta}. \end{cases}$$

Part (c) and (d) come from the following best-response arguments. Our goal is to show $p^*_{\text{signal aware}}$ is a best response given q^* and vice versa, where

$$p^*_{\text{signal aware}}(s) = \begin{cases} \underline{\theta} & \text{if } s = \underline{s} \text{ is observed} \\ \overline{\theta} & \text{if signal } s = \overline{s} \text{ is observed.} \end{cases}$$

and

$$q^* = \min\left\{1, \frac{(1-\mu)\underline{\theta}}{\mu\Delta\theta}\right\}$$

What is the signal aware seller's best response after seeing $\overline{s}$*?* From part (a), we know that $\underline{\theta}$ buyers never signal $\overline{\theta}$, so the seller knows that a $\overline{s}$ signal implies that the buyer is type $\overline{\theta}$ and should therefore set a price of $\overline{\theta}$ after seeing $\overline{s}$, i.e., $p^*_{\text{signal aware}}(\overline{s}) = \overline{\theta}$.

What is the signal aware seller's best response after seeing $\underline{s}$*?* In order for $p^*_{\text{signal aware}}$ to be a best response, it must maximize the seller's expected utility, where the expectation is over the seller's posterior belief over the buyer's type given that they have signaled $\underline{s}$. Given probability q^* that the $\overline{\theta}$ buyer sends signal $\underline{s}$, the seller's posterior belief $\hat{\mu}$ that the buyer is type $\overline{\theta}$ is

$$\hat{\mu} = \mathbb{P}(\theta = \overline{\theta}|s = \underline{s}) = \frac{\mathbb{P}(s = \underline{s}|\theta = \overline{\theta})\mathbb{P}(\theta = \overline{\theta})}{\mathbb{P}(s = \underline{s}|\theta = \overline{\theta})\mathbb{P}(\theta = \overline{\theta}) + \mathbb{P}(s = \underline{s}|\theta = \underline{\theta})\mathbb{P}(\theta = \underline{\theta})} = \frac{q^*\mu}{q^*\mu + 1 - \mu}.$$

Let $f(p)$ denote the seller's expected utility from charging price p after observing signal $\underline{s}$, so

$$f(p) = \begin{cases} p - \hat{\mu}q^*c_S & \text{if } p < \underline{\theta} \\ \hat{\mu}p - \hat{\mu}q^*c_S & \text{if } p \in [\underline{\theta}, \overline{\theta}]. \end{cases}$$

In order for $p^*_{\text{signal aware}}(\underline{s})$ to be a best response, it must be the value that maximizes f:

$$p^*_{\text{signal aware}}(\underline{s}) = \max_p f(p) = \begin{cases} \underline{\theta} & \text{if } q^* \leq \min\left\{1, \frac{(1-\mu)\underline{\theta}}{\mu\Delta\theta}\right\} \\ \overline{\theta} & \text{else.} \end{cases} = \underline{\theta},$$

where the last equality comes from the choice of q^*. This shows that $p^*_{\text{signal aware}}(\underline{s}) = \underline{\theta}$ is a best response for the seller. We now turn our attention to the $\overline{\theta}$-buyer.

What is the optimal probability q^* of evasion for the $\overline{\theta}$-buyer? Let $g(q)$ denote the the expected utility for the $\overline{\theta}$ buyer when they evade with probability q, given that the seller is playing $p^*_{signalblind}$ if they are signal blind and $p^*_{signalaware}$ if they are signal aware, so

$$\begin{aligned}
g(q) &= \mathbb{P}(\text{seller is signal blind})(\overline{\theta}\text{-buyer utility if seller plays } p^*_{\text{signal blind}}) \\
&\qquad + \mathbb{P}(\text{seller is signal aware})(\overline{\theta}\text{-buyer utility if seller plays } p^*_{\text{signal aware}}) \\
&= (1-\alpha)(\overline{\theta}\text{-buyer utility if seller plays } p^*_{\text{signal blind}}) \\
&\qquad + \alpha(\overline{\theta}\text{-buyer utility if seller plays } p^*_{\text{signal aware}}) \\
&= (1-\alpha)[(\Delta\theta - c_B q)\mathbb{1}(\underline{\theta} \geq \mu\overline{\theta}) + (-c_B q)\mathbb{1}(\underline{\theta} < \mu\overline{\theta})] \\
&\qquad + \alpha[(\Delta\theta - c_B)q\mathbb{1}(q \leq \min\{1, {}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta}\}) + (-c_B q)\mathbb{1}(q > \min\{1, {}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta}\})] \qquad (4.4)
\end{aligned}$$

We analyze (4.4) in cases.

- If $1 \leq {}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta}$, this implies that $\underline{\theta} \geq \mu\overline{\theta}$, so (4.4) simplifies to

$$u_B = (1-\alpha)\Delta\theta + (\alpha\Delta\theta - c_B)q.$$

- If ${}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta} \leq 1$, this implies $\theta < \mu\overline{\theta}$, so (4.4) simplifies to

$$u_B = \begin{cases} (\alpha\Delta\theta - c_B)q & \text{if } q \leq {}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta} \\ -c_B q & \text{else.} \end{cases}$$

Combining both cases, we see that the $\overline{\theta}$-buyer's optimal probability of evasion is

$$q^* = \begin{cases} \min\left\{1, \frac{(1-\mu)\underline{\theta}}{\mu\Delta\theta}\right\} & \text{if } \alpha > {}^{c_B}/_{\Delta\theta} \\ 0 & \text{else.} \end{cases}$$

□

Proof of Corollary 2

Proof. We summarize fundamental properties of the equilibrium in Theorem 6, from which expressions for the buyer's and seller's expected utilities follow. At equilibrium,

1. When $\underline{\theta} \geq \mu\overline{\theta}$, the signal blind seller always sets price $\underline{\theta}$.

2. When $\underline{\theta} < \mu\overline{\theta}$, the signal blind seller always sets price $\overline{\theta}$.

3. When $\underline{\theta} \geq \mu\overline{\theta}$ and $\alpha > {}^{c_B}/_{\Delta\theta}$, the $\overline{\theta}$-buyer always evades (since $1 \leq {}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta}$ when $\underline{\theta} \geq \mu\overline{\theta}$).

4. When $\underline{\theta} < \mu\overline{\theta}$ and $\alpha > {}^{c_B}/_{\Delta\theta}$, the $\overline{\theta}$-buyer evades with probability ${}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta}$ (since ${}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta} < 1$ when $\underline{\theta} < \mu\overline{\theta}$).
5. The $\overline{\theta}$-buyer always signals truthfully when $\alpha \leq {}^{c_B}/_{\Delta\theta}$.
6. The $\underline{\theta}$-buyer always signals truthfully.

Buyer Utilities. It is straightforward to see that the $\underline{\theta}$-buyer's expected utility is zero, so we focus on the $\overline{\theta}$-buyer. The $\overline{\theta}$-buyer's expected utility is

$$\begin{aligned} u_B &= \mathbb{P}(\text{seller is signal blind})[u_B|\text{seller is signal blind}] \\ &\quad + \mathbb{P}(\text{seller is signal aware})[u_B|\text{seller is signal aware}] \\ &= (1-\alpha)[u_B|\text{seller is signal blind}] + \alpha[u_B|\text{seller is signal aware}]. \end{aligned} \tag{4.5}$$

where, by the properties above,

- If $\underline{\theta} \geq \mu\overline{\theta}$,

$$\begin{aligned} u_B|\text{seller is signal blind} &= (\Delta\theta - c_B)\mathbb{1}(\alpha > {}^{c_B}/_{\Delta\theta}) + \Delta\theta\mathbb{1}(\alpha \leq {}^{c_B}/_{\Delta\theta}). \\ u_B|\text{seller is signal aware} &= (\Delta\theta - c_B)\mathbb{1}(\alpha \leq {}^{c_B}/_{\Delta\theta}). \end{aligned}$$

- If $\underline{\theta} < \mu\overline{\theta}$,

$$\begin{aligned} u_B|\text{seller is signal blind} &= -c_B({}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta})\mathbb{1}(\alpha > {}^{c_B}/_{\Delta\theta}) \\ u_B|\text{seller is signal aware} &= (\Delta\theta - c_B)({}^{(1-\mu)\underline{\theta}}/_{\mu\Delta\theta})\mathbb{1}(\alpha \leq {}^{c_B}/_{\Delta\theta}) \end{aligned}$$

Seller Utility. The seller's expected utility is

$$\begin{aligned} u_S &= \mathbb{P}(\text{seller is signal blind})[u_S|\text{seller is signal blind}] \\ &\quad + \mathbb{P}(\text{seller is signal aware})[u_S|\text{seller is signal aware}] \\ &= (1-\alpha)[u_S|\text{seller is signal blind}] + \alpha[u_S|\text{seller is signal aware}]. \end{aligned} \tag{4.6}$$

where, by the properties above,

- If $\underline{\theta} \geq \mu\overline{\theta}$,

$$\begin{aligned} &u_S|\text{seller is signal blind} \\ &= \mathbb{P}(\text{buyer is type-}\underline{\theta})(u_S|\text{seller is signal blind and buyer is type-}\underline{\theta}) \\ &\quad + \mathbb{P}(\text{buyer is type-}\overline{\theta})(u_S|\text{seller is signal blind and buyer is type-}\overline{\theta}) \\ &= (1-\mu)\underline{\theta} + \mu[(\underline{\theta} - c_S)\mathbb{1}(\alpha > {}^{c_B}/_{\Delta\theta}) + \underline{\theta}\mathbb{1}(\alpha \leq {}^{c_B}/_{\Delta\theta})] \end{aligned}$$

and

$$\begin{aligned}
&u_S|\text{seller is signal aware}\\
&= \mathbb{P}(\text{buyer is type-}\underline{\theta})(u_S|\text{seller is signal aware and buyer is type-}\underline{\theta})\\
&\quad + \mathbb{P}(\text{buyer is type-}\overline{\theta})(u_S|\text{seller is signal aware and buyer is type-}\overline{\theta})\\
&= (1-\mu)\underline{\theta} + \mu[(\underline{\theta} - c_S)\mathbb{1}(\alpha > {}^{c_B}\!/\!_{\Delta\theta}) + \overline{\theta}\mathbb{1}(\alpha \le {}^{c_B}\!/\!_{\Delta\theta})].
\end{aligned}$$

- If $\underline{\theta} < \mu\overline{\theta}$,

$$\begin{aligned}
&u_S|\text{seller is signal blind}\\
&= \mathbb{P}(\text{buyer is type-}\underline{\theta})(u_S|\text{seller is signal blind and buyer is type-}\underline{\theta})\\
&\quad + \mathbb{P}(\text{buyer is type-}\overline{\theta})(u_S|\text{seller is signal blind and buyer is type-}\overline{\theta})\\
&= (1-\mu)0\\
&\quad + \mu[(\overline{\theta} - c_S)({}^{(1-\mu)\underline{\theta}}\!/\!_{\mu\Delta\theta})\mathbb{1}(\alpha > {}^{c_B}\!/\!_{\Delta\theta})\\
&\quad + \overline{\theta}(1 - ({}^{(1-\mu)\underline{\theta}}\!/\!_{\mu\Delta\theta}))\mathbb{1}(\alpha > {}^{c_B}\!/\!_{\Delta\theta}) + \overline{\theta}\mathbb{1}(\alpha \le {}^{c_B}\!/\!_{\Delta\theta})]
\end{aligned}$$

and

$$\begin{aligned}
&u_S|\text{seller is signal aware}\\
&= \mathbb{P}(\text{buyer is type-}\underline{\theta})(u_S|\text{seller is signal aware and buyer is type-}\underline{\theta})\\
&\quad + \mathbb{P}(\text{buyer is type-}\overline{\theta})(u_S|\text{seller is signal aware and buyer is type-}\overline{\theta})\\
&= (1-\mu)\underline{\theta}\\
&\quad + \mu[(\underline{\theta} - c_S)({}^{(1-\mu)\underline{\theta}}\!/\!_{\mu\Delta\theta})\mathbb{1}(\alpha > {}^{c_B}\!/\!_{\Delta\theta})\\
&\quad + \overline{\theta}(1 - ({}^{(1-\mu)\underline{\theta}}\!/\!_{\mu\Delta\theta}))\mathbb{1}(\alpha > {}^{c_B}\!/\!_{\Delta\theta}) + \overline{\theta}\mathbb{1}(\alpha \le {}^{c_B}\!/\!_{\Delta\theta})].
\end{aligned}$$

Plugging in the relevant values of α to the buyer's (4.5) and seller's (4.6) utility expressions gives the stated orderings. We also see that $\alpha^* = {}^{c_B}\!/\!_{\Delta\theta}$ maximizes the seller's utility. □

Proof of Proposition 6

Proof. For each round t, let

$$I_t = \mathbb{1}\{\exists i \text{ s.t. } s_i^t = \overline{s} \text{ and } \exists j \text{ s.t. } s_j^t = \underline{s}\}$$

be an indicator for whether both types of signals are observed at round t, i.e., whether round t is "informative" about if there is price discrimination. For rounds t with $I_t = 1$, we additionally define the following random variables:

- $\overline{P}_t = p_i^t$ for the smallest $i \in [N]$ such that $s_i^t = \overline{s}$,

- $\underline{P}_t = p_j^t$ for the smallest $j \in [N]$ such that $s_j^t = \underline{s}$, and
- $X_t = \mathbb{1}\{\overline{P}_t \neq \underline{P}_t\}$, an indicator for observed price discrimination.

Note that the choice to define $\overline{P}_t$ and $\underline{P}_t$ to correspond to the *smallest* index satisfying the corresponding condition is simply for concreteness; we could equivalently sample uniformly from the set of indices satisfying the condition.

Recall that $H_t = ((\theta_i^\tau, s_i^\tau, p_i^\tau)_{i=1}^n)_{\tau=1}^{t-1}$ is the history known by buyers at the beginning of round t. Consider the following estimator:

$$\hat{\alpha}_t = \frac{1}{t}\sum_{\tau=1}^{t-1} \frac{X_\tau I_\tau}{\mathbb{E}[I_\tau | H_\tau]} \tag{4.7}$$

The expectation $\mathbb{E}[I_\tau|H_\tau]$ is over the randomness at round τ. Note that $\hat{\alpha}_t$ is computable based on the history H_t, because $\mathbb{E}[I_\tau|H_\tau]$ is computable for any $\tau < t$. We will now show that $\hat{\alpha}_t$ satisfies Definition 6. We start by computing the expectation of $\hat{\alpha}_t$:

$$\begin{aligned}
\mathbb{E}[\hat{\alpha}_t] &= \mathbb{E}\left[\frac{1}{t}\sum_{\tau=1}^{t-1} \frac{X_\tau I_\tau}{\mathbb{E}[I_\tau|H_\tau]}\right] \\
&= \frac{1}{t}\sum_{\tau=1}^{t-1} \mathbb{E}\left[\frac{X_\tau I_\tau}{\mathbb{E}[I_\tau|H_\tau]}\right] && \text{linearity of expectation} \\
&= \frac{1}{t}\sum_{\tau=1}^{t-1} \mathbb{E}\left[\mathbb{E}\left[\frac{X_\tau I_\tau}{\mathbb{E}[I_\tau|H_\tau]}\middle| H_\tau\right]\right] && \text{tower rule} \\
&= \frac{1}{t}\sum_{\tau=1}^{t-1} \mathbb{E}\left[\frac{\mathbb{E}[X_\tau I_\tau|H_\tau]}{\mathbb{E}[I_\tau|H_\tau]}\right]
\end{aligned}$$

Observe that X_τ and I_τ are independent given H_τ. To see why, note that the randomness in $X_\tau|H_\tau$ comes only from the randomness in the seller's mixed strategy at round τ, whereas the randomness in $I_\tau|H_\tau$ comes only from the randomness in the buyers mixed strategy at round τ. The mixed strategies are fixed given H_τ, and the additional randomness is independent. Thus,

$$\begin{aligned}
&= \frac{1}{t}\sum_{\tau=1}^{t-1} \mathbb{E}\left[\frac{\mathbb{E}[X_\tau|H_\tau]\mathbb{E}[I_\tau|H_\tau]}{\mathbb{E}[I_\tau|H_\tau]}\right] \\
&= \frac{1}{t}\sum_{\tau=1}^{t-1} \mathbb{E}[\mathbb{E}[X_\tau|H_\tau]] \\
&= \frac{1}{t}\sum_{\tau=1}^{t-1} \mathbb{E}\left[\mathbb{E}[\mathbb{1}\{\overline{P}_\tau \neq \underline{P}_\tau\}|H_\tau]\right] && \text{by definition of } X_\tau
\end{aligned}$$

Since $\overline{P}_\tau | H_\tau \sim \pi_t^p(\cdot | s = \overline{s})$ and $\underline{P}_\tau | H_\tau \sim \pi_t^p(\cdot | s = \underline{s})$ by definition of the game, we have

$$= \frac{1}{t} \sum_{\tau=1}^{t-1} \alpha_\tau$$

Finally, plugging in the above expression with $t = T$ into the criterion for consistency, we have

$$\lim_{T \to \infty} \left| \mathbb{E}[\hat{\alpha}_T] - \frac{1}{T} \sum_{t=1}^{T} \alpha_t \right| = \lim_{T \to \infty} \left| \frac{1}{T} \sum_{t=1}^{T-1} \alpha_t - \frac{1}{T} \sum_{t=1}^{T} \alpha_t \right| = \lim_{T \to \infty} \frac{\alpha_T}{T} = 0$$

as desired. The last equality comes from the fact that α_T is a probability, so it is bounded between 0 and 1 for all T. □

Proof of Proposition 7

Proof. First, we will show that always price-discriminating ($\pi_t^p = \pi_{\text{PD}}^p$ for all $t \in [T]$) is no-regret against CBER-buyers. For CBER-buyers, their strategy π_t^s at each round t is either π_{truthful}^s or $\pi_{\text{strategic}}^s$. For both these buyer responses, the seller's optimal strategy is to always price discriminate as shown in the computation of the seller's equilibrium response in the proof of Theorem 6. In other words, the seller incurs zero regret in each round and thus zero average regret.

Next, we will analyze the seller's average utility. Note that when $\pi_t^p = \pi_{\text{PD}}^p$, the probability of seeing different prices for different signals is $\alpha_t = 1$, so $(1/t) \sum_{\tau=1}^{t} \alpha_\tau = 1$ for all t. By the consistency property, $\hat{\alpha}_t$ becomes greater than α^* eventually (where α^* is as defined in Corollary 3) and the buyer plays $\pi_{\text{strategic}}^s$. In other words, eventually the seller and buyers will all be playing their equilibrium strategies for the PD-game with $\alpha = 1$, so their average utilities will converge to the corresponding equilibrium utilities. We make this argument formal below.

Define $\kappa < \infty$ to be the maximum utility that can be achieved by a seller in any round. The finiteness of κ is guaranteed by definition of the seller's utility function. Define $A_T = \{\exists t > \sqrt{T} \text{ s.t. } \hat{\alpha}_t > \alpha^*\}$ and let $A_T^C = \{\hat{\alpha}_t > \alpha^* \text{ for all } t > \sqrt{T}\}$ denote the complement. Let $\gamma_T = \mathbb{P}(A_T)$ and $1 - \gamma_T = \mathbb{P}(A_T^C)$ denote the corresponding probabilities. Then, we can decompose the expected average seller's utility as

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} U_S(\pi_t)\right] = \gamma_t \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} U_S(\pi_t) \middle| A_T\right] + (1 - \gamma_t) \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} U_S(\pi_t) \middle| A_T^C\right]. \quad (4.8)$$

The first term of (4.8) is trivially upper bounded by $\gamma_T \kappa$.

To bound the second term of (4.8), first note that for any round t where $\hat{\alpha}_t > \alpha^*$, the buyer's strategy will be equivalent to their equilibrium strategy with $\alpha = 1$. Thus, the best utility that the seller can achieve for those rounds is $u_S(1)$. It follows that under the condition that $\hat{\alpha}_t > \alpha^*$ for every $t > \sqrt{T}$, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t) &= \frac{1}{T}\sum_{t=\sqrt{T}}^{T} U_S(\pi_t) + \frac{1}{T}\sum_{t=1}^{\sqrt{T}} U_S(\pi_t) \\
&\leq \frac{1}{T}\sum_{t=\sqrt{T}}^{T} u_S(1) + \frac{1}{T}\sum_{t=1}^{\sqrt{T}} \kappa \\
&= \frac{T-\sqrt{T}}{T} u_S(1) + \frac{\sqrt{T}\kappa}{T} \\
&\leq u_S(1) + \frac{\kappa - u_S(1)}{\sqrt{T}}.
\end{aligned}$$

Plugging back into (4.8), we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t)\right] \leq \gamma_t\kappa + (1-\gamma_t)\left(u_S(1) + \frac{\kappa - u_S(1)}{\sqrt{T}}\right).$$

By the consistency property (Lemma 26), we know $\lim_{T\to\infty}\gamma_T = 0$, so

$$\lim_{T\to\infty}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t)\right] \leq u_S(1).$$

□

Missing Proofs of Lemmas in Proof of Proposition 8

Proof of Lemma 27. Based on the utility orderings from Corollary 2, note the following ordering of seller utilities for different combinations of buyer and seller policies:

$$U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{PD}}) > U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{noPD}}) > U_S(\pi^s_{\text{strategic}}, \pi^p_{\text{PD}}) > U_S(\pi^s_{\text{strategic}}, \pi^p),$$

where π^p is any other pricing strategy besides π^p_{PD} and π^p_{noPD}. We can then write

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t) &\leq \frac{1}{T}\sum_{t\in\mathcal{T}} U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{noPD}}) + \sum_{t\in[T]\setminus\mathcal{T}} U_S(\pi^s_{\text{strategic}}, \pi^p_{\text{PD}}) \\
&= \frac{|\mathcal{T}|}{T} U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{noPD}}) + \left(1 - \frac{|\mathcal{T}|}{T}\right) U_S(\pi^s_{\text{strategic}}, \pi^p_{\text{PD}})
\end{aligned}$$

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t) \leq U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{noPD}}) \lim_{T\to\infty} \frac{|\mathcal{T}|}{T} + U_S(\pi^s_{\text{strategic}}, \pi^p_{\text{PD}}) \left(1 - \lim_{T\to\infty} \frac{|\mathcal{T}|}{T}\right)$$

Since $U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{PD}}) > U_S(\pi^s_{\text{strategic}}, \pi^p_{\text{PD}})$, the above upper bound on the limit of the average utility is increasing as $\lim_{T\to\infty} |\mathcal{T}|/T$ is increasing. When $\lim_{T\to\infty} |\mathcal{T}|/T \leq \alpha^*$,

$$\begin{aligned}
&\leq \alpha^* U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{PD}}) + (1-\alpha^*) U_S(\pi^s_{\text{strategic}}, \pi^p_{\text{PD}}) \\
&< \alpha^* U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{PD}}) + (1-\alpha^*) U_S(\pi^s_{\text{truthful}}, \pi^p_{\text{PD}}) \\
&= \mathbb{U}^*_1
\end{aligned}$$

□

Proof of Lemma 28. Let R^S_T denote the average seller utility in the T rounds. Since the seller is no regret, $\lim_{T\to\infty} R^S_T = 0$.

Consider the regret due to the seller deviating to π^p_{PD} in each round. The gain in utility in each round due to this deviation is non-negative since π^p_{PD} is the best-response to both possible buyer strategies $\pi^s_{\text{truthful}}, \pi^s_{\text{strategic}}$. We can then lower bound the regret by considering regret accumulated in rounds where $\hat{\alpha}_t \leq \alpha^*$. In such rounds, all buyers are truthful, so whenever the seller does not charge a buyer the price corresponding to their signal type, they incur regret. The probability that the seller observes $\overline{s}$ but charges $\overline{\theta}$ is $\mu\pi^p_t(\underline{\theta}|\underline{s})$, and this yields a loss of utility of $\Delta\theta$, because the buyer is type $\overline{\theta}$. Similarly, the probability that the seller observes $\underline{s}$ but charges $\overline{\theta}$ is $(1-\mu)\pi^p_t(\overline{\theta}|\underline{s})$, and this yields a loss of utility of $\underline{\theta}$, since the buyer is type $\underline{\theta}$.

$$R^S_T \geq \frac{1}{T} \sum_{t:\hat{\alpha}_t \leq \alpha^*} \mu\Delta\theta\pi^p_t(\underline{\theta}|\overline{s}) + (1-\mu)\underline{\theta}\pi^p_t(\overline{\theta}|\underline{s})$$

Let $\kappa = \min\{\mu\Delta\theta, (1-\mu)\underline{\theta}\}$

$$\geq \frac{1}{T} \sum_{t:\hat{\alpha}_t \leq \alpha^*} \kappa(1 - (\pi^p(\overline{\theta}|\overline{s}) - \pi^p(\overline{\theta}|\underline{s})))$$

$$\implies \qquad \frac{1}{T}\sum_{t\in\mathcal{T}} \left(\pi^p_t(\overline{\theta}|\overline{s}) - \pi^p_t(\overline{\theta}|\underline{s})\right) \geq \frac{|\mathcal{T}|}{T} - \frac{R^S_T}{\kappa}$$

The above inequality shows that $|\mathcal{T}|/T$ is bounded above by some measure of simultaneous truthfulness and price discrimination. Each quantity $\pi^p_t(\overline{\theta}|\overline{s}) - \pi^p_t(\overline{\theta}|\overline{s})$ is a measure of price-discrimination in each round and is related to α_t as described in the following lemma.

Lemma 30. *When seller pricing strategies are supported on* $\{\overline{\theta}, \underline{\theta}\}$, $\alpha_t \geq \pi^p_t(\overline{\theta}|\overline{s}) - \pi^p_t(\overline{\theta}|\overline{s})$

Proof of Lemma 30. Since seller pricing strategies are supported on $\{\overline{\theta}, \underline{\theta}\}$, α_t which is the probability of seeing different prices for different signals is

$$\begin{aligned}\alpha_t &= \pi_t^p(\overline{\theta}|\overline{s})\pi_t^p(\underline{\theta}|\underline{s}) + \pi_t^p(\underline{\theta}|\overline{s})\pi_t^p(\overline{\theta}|\underline{s})\\ &= \pi_t^p(\overline{\theta}|\overline{s})(1-\pi_t^p(\overline{\theta}|\underline{s})) + (1-\pi_t^p(\overline{\theta}|\overline{s}))\pi_t^p(\overline{\theta}|\underline{s})\\ &= \pi_t^p(\overline{\theta}|\overline{s}) + \pi_t^p(\overline{\theta}|\underline{s}) - 2\pi_t^p(\overline{\theta}|\overline{s})\pi_t^p(\overline{\theta}|\underline{s})\\ &= \pi_t^p(\overline{\theta}|\overline{s}) - \pi_t^p(\overline{\theta}|\underline{s}) + 2\pi_t^p(\overline{\theta}|\underline{s})(1-\pi_t^p(\overline{\theta}|\overline{s}))\\ &\geq \pi_t^p(\overline{\theta}|\overline{s}) - \pi_t^p(\overline{\theta}|\underline{s})\end{aligned}$$

□

By inequality 1, Lemma 30, and since $\lim_{T\to\infty} R_T^S/\kappa = 0$,

$$\lim_{T\to\infty}\frac{|\mathcal{T}|}{T} \leq \lim_{T\to\infty}\frac{1}{T}\sum_{t\in\mathcal{T}}\alpha_t.$$

□

Proof of Lemma 29. Consider the last index t^* in $\mathcal{T}$. Let us consider two cases. The first case is $\lim_{T\to\infty} t^*/T < \alpha^*$. Then,

$$\begin{aligned}\sum_{t\in\mathcal{T}}\alpha_t/T &\leq |\mathcal{T}|/T\\ &\leq t^*/T\\ \implies \lim_{T\to\infty}\sum_{t\in\mathcal{T}}\alpha_t/T &\leq \alpha^*.\end{aligned}$$

In the second case, $\lim_{T\to\infty} t^* = \infty$. Consider $\bar{\alpha}_{t^*} = \frac{1}{t^*}\sum_{t\leq t^*}\alpha_t \geq \sum_{t\in\mathcal{T}}\alpha_t/T$. By the consistency property, $\lim_{T\to\infty}\bar{\alpha}_{t^*} = \hat{\alpha}_{t^*}$. $\hat{\alpha}_{t^*} \leq \alpha^*$ since $t^* \in \mathcal{T}$. □

Proof of Proposition 9

Proof. Under the conditions of this proposition, the seller's utility must, by definition of policy regret, approach a utility at least as high (or better) than the utility of any strategy in $\mathbb{A}_S^{MS}$ as $T\to\infty$. Recall that $\mathbb{U}_1^*$ is the seller utility achieved in the PD game when $\alpha = \alpha^*$. For any $\epsilon > 0$, there exists a value $\tilde{\alpha} < \alpha^*$ such that the seller's utility in the PD game with $\alpha = \tilde{\alpha}$ (denoted $u_S(\tilde{\alpha})$) is at least $\mathbb{U}_1^* - \epsilon$. Consider the seller's mixed strategy of price-discriminating with probability $\tilde{\alpha}$, so $\alpha_t = \tilde{\alpha}$ for all t. Similar to the proof of Proposition 7, define $A_T = \{\exists t > \sqrt{T} \text{ s.t. } \hat{\alpha}_t > \alpha^*\}$ and let $A_T^C = \{\hat{\alpha}_t > \alpha^* \text{ for all } t > \sqrt{T}\}$ denote the complement. Let $\gamma_T = \mathbb{P}(A_T)$ and $1-\gamma_T = \mathbb{P}(A_T^C)$ denote the corresponding probabilities. The expected average seller's utility can be decomposed as $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t)\right] = \gamma_t \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t)|A_T\right] + (1-\gamma_t)\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t)|A_T^C\right]$. Define m to be the smallest utility that can be achieved by a seller in any round. Then the first term on the right side is trivially lower bounded by $\gamma_T m$.

To bound the second term, note that for any round t where $\hat{\alpha}_t < \alpha^*$, the buyer's strategy will be equivalent to the equilibrium strategy of the PD game with $\alpha = \tilde{\alpha}$, so the seller's expected utility is $u_S(\tilde{\alpha})$. Thus, $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t)|A_T^C\right] = \frac{1}{T}\sum_{t=\sqrt{T}}^{T} \mathbb{E}[U_S(\pi_t) \mid \alpha_t < \alpha^*] + \frac{1}{T}\sum_{t=1}^{\sqrt{T}} \mathbb{E}[U_S(\pi_t)\alpha_t < \alpha^*] \geq \frac{1}{T}\sum_{t=\sqrt{T}}^{T} u_S(\tilde{\alpha}) + \frac{1}{T}\sum_{t=1}^{\sqrt{T}} m = u_S(\tilde{\alpha}) + (m - u_s(\tilde{\alpha}))/\sqrt{T}$

Plugging back into the expected average seller's utility yields $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t)\right] \geq \gamma_T m + (1-\gamma_T)[u_s(\tilde{\alpha}) + (m - u_s(\tilde{\alpha}))/\sqrt{T}]$. Since the seller is playing $\alpha_t \equiv \tilde{\alpha} < \alpha^*$, the consistency of the buyer's beliefs tells us that $\lim_{T\to\infty} \gamma_T = 0$, so $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} U_S(\pi_t)\right] \geq u_S(\tilde{\alpha}) = \mathbb{U}_1^* - \epsilon$. Taking ϵ to 0 gives the desired result. □

Proof of Proposition 10

Proof. Similar to the proof of Proposition 8, we will show that due to the consistency property of the beliefs $(\hat{\alpha}_t)$, we cannot simultaneously have a high degree of price-discrimination and truthful behavior from buyers. And this will imply that the seller cannot be better than $\mathbb{U}_1^*$ asymptotically.

We will provide the proof for the case $\underline{\theta} \leq \mu\overline{\theta}$ and the proof for the other case follows similarly.

Let us compare the cumulative seller utilities due to a sequence of $(\pi_t)_{t=1}^T$ versus $T \cdot \mathbb{U}_1^*$. Let us denote by π^{p*} the commitment strategy given in Corollary 3. Then $\mathbb{U}_1^* = U_S(\pi^s_{\text{truthful}}, \pi^{p*})$. In the case of $\underline{\theta} \leq \mu\overline{\theta}$, this is the strategy where $\pi^p(\underline{\theta}|\underline{s}) = 1$ and $\pi^p(\overline{\theta}|\overline{s}) = \alpha^*$.
$\sum_{t=1}^{T}(U_S(\pi_t) - U_S(\pi^*))$ is

$$\sum_{t\in\mathcal{T}} (U_S(\pi^s_{\text{truthful}}, \pi^p_t) - U_S(\pi^s_{\text{truthful}}, \pi^{p*})) + \sum_{t\notin\mathcal{T}} \left(U_S(\pi^s_{\text{strategic}}, \pi^p_t) - U_S(\pi^s_{\text{truthful}}, \pi^{p*})\right)$$

Note that for all π^p, $U_S(\pi^s_{\text{strategic}}, \pi^p_t) \leq U_S(\pi^s_{\text{strategic}}, \pi^p_{\text{PD}}) < U_S(\pi^s_{\text{truthful}}, \pi^{p*})$.

$$\begin{aligned} U_S(\pi^s_{\text{truthful}}, \pi^{p*}) - U_S(\pi^s_{\text{strategic}}, \pi^p_{\text{PD}}) &= \alpha^*\mu\overline{\theta} + \alpha^*(1-\mu)\underline{\theta} + (1-\alpha^*)\underline{\theta} - \underline{\theta} + \mu q^* c_S \\ &\geq \alpha^*\mu\Delta\theta. \end{aligned}$$

$$\sum_{t=1}^{T}(U_S(\pi_t) - U_S(\pi^*)) \leq \sum_{t\in\mathcal{T}} [U_S(\pi^s_{\text{truthful}}, \pi^p_t) - U_S(\pi^s_{\text{truthful}}, \pi^{p*})]$$

$$- \alpha^*(T - |\mathcal{T}|)\mu\Delta\theta$$

Note that for any π^p, the seller's utility when buyers signal truthfully is

$$U_S(\pi^s_{\text{truthful}}, \pi^p) = \mu\overline{\theta}\pi^p(\overline{\theta}|\overline{s}) + \mu\underline{\theta}\pi^p(\underline{\theta}|\overline{s}) + (1-\mu)\underline{\theta}\pi^p(\underline{\theta}|\underline{s}) + 0 \cdot \pi^p(\overline{\theta}|\underline{s})$$

Since $\underline{\theta} \leq \mu\overline{\theta}$, $\pi^{p^*}(\underline{\theta}|\underline{s}) = 1$ and $\pi^{p^*}(\overline{\theta}|\overline{s}) = \alpha^*$.

$$\begin{aligned} U_S(\pi^s_{\text{truthful}}, \pi^p) - U_S(\pi^s_{\text{truthful}}, \pi^{p^*}) &= \mu\overline{\theta}(\pi^p\overline{\theta}|\overline{s}) - \alpha^*) + \mu\underline{\theta}(1 - \pi^p - (1-\alpha^*)) \\ &\quad + (1-\mu)\underline{\theta}(\pi^p(\underline{\theta}|\underline{s}) - 1) \\ &= \mu\Delta\theta(\pi^p\overline{\theta}|\overline{s}) - \alpha^*) + (1-\mu)\underline{\theta}(\pi^p(\underline{\theta}|\underline{s}) - 1) \end{aligned}$$

$$\begin{aligned} \sum_{t=1}^{T}(U_S(\pi_t) - U_S(\pi^*)) &\leq \sum_{t\in\mathcal{T}} \left[\mu\Delta\theta(\pi^p(\overline{\theta}|\overline{s}) - \alpha^*) - (1-\mu)\underline{\theta}(1 - \pi^p(\underline{\theta}|\underline{s}))\right] \\ &\quad - \alpha^*(T - |\mathcal{T}|)\mu\Delta\theta \end{aligned}$$

Since $\mu\overline{\theta} < \underline{\theta}$, $(1-\mu)\underline{\theta} > \mu\Delta\theta$. So,

$$\begin{aligned} &< \sum_{t\in\mathcal{T}} \mu\Delta\theta(\pi^p_t(\overline{\theta}|\overline{s}) - \pi^p_t(\underline{\theta}|\underline{s}) - \alpha^*) - \alpha^*(T - |\mathcal{T}|)\mu\Delta\theta \\ &= \sum_{t\in\mathcal{T}} \mu\Delta\theta(\pi^p_t(\overline{\theta}|\overline{s}) - \pi^p_t(\underline{\theta}|\underline{s})) - \alpha^* T\mu\Delta\theta \\ &\leq \mu\Delta\theta \sum_{t\in\mathcal{T}} \alpha_t - \alpha^* T\mu\Delta\theta \quad \text{(By Lemma 30)} \\ \implies \lim_{T\to\infty} \sum_{t=1}^{T}(U_S(\pi_t) - U_S(\pi^*)) &\leq \mu\Delta\theta\left(\lim_{T\to\infty}\frac{1}{T}\sum_{t\in\mathcal{T}}\alpha_t - \alpha^*\right) \\ &\leq 0 \quad \text{(By Lemma 29)} \end{aligned}$$

□

Chapter 5

Conclusion

In order to leverage the full power of collaborative learning, the incentives of the participants must be accounted for – otherwise they are at risk of defecting from the system. This dissertation examines three topics at the intersection of incentives and collaborative learning. I summarize the contributions herein and propose directions for future work.

I. Personalized Collaborative Learning. In a collaborative learning system, when a global model is trained on the clients' aggregate data, it may perform poorly on individual client tasks. Identifying clients with similar objectives and learning a model-per-cluster is an intuitive and interpretable approach to this personalization problem. However, doing so with provable and optimal guarantees has remained an open challenge. We formalize this problem as a stochastic optimization problem, achieving optimal convergence rates for a large class of loss functions. We propose simple iterative algorithms which identify clusters of similar clients and train a personalized model-per-cluster, using local client gradients and flexible constraints on the clusters. The convergence rates of our algorithms asymptotically match those obtained if we knew the true underlying clustering of the clients and are provably robust in the Byzantine setting where some fraction of the clients are malicious.

Future Directions. Theorem 1 gives an upper bound of $\mathcal{O}(\sigma^3/\Delta)$ on the error of our clustering algorithm, while Theorem 2 establishes a lower bound on any cluster-identification algorithm for our problem setting of $\mathcal{O}(\sigma^4/\Delta^2)$. If Δ is large, the gap between the lower and upper bound may also be large. Designing an algorithm, or modifying the current analysis, to close this gap is a direction for future work. Additionally, deriving rates as in Theorem 3 for more structured losses (e.g. convex, strongly-convex) remains to be done.

II. Collaborative Learning among Competitors. We study collaborative learning systems in which the participants are competitors who will defect from the system if they lose revenue by collaborating. As such, we frame the system as a duopoly of competitive firms who are each training machine learning models and selling their predictions to a market of consumers. We first examine a fully collaborative scheme in which both firms share their models with each other and show that this leads to a market collapse with the revenues of both firms going to zero. We next show that one-sided collaboration in which only the firm with the lower-quality model shares improves the revenue of both firms. Finally, we propose a more equitable, *defection-free* scheme in which both firms share with each other while losing no revenue. We show for all but trivial starting conditions our algorithm converges to the Nash bargaining solution.

Future Directions. There is room to expand our theory to a more general problem setting. For instance we study a duopoly, not a general oligopoly, we assume consumer types are uniformly distributed, we derive convergence rates only for convex losses, and we only consider the non-stochastic setting. Deriving all these results in more generality is open for future work, in addition to a more rigorous analysis of consumer utility.

III. Privacy Dynamics in Systems of Learning Agents. In many real-world settings, if clients in a collaborative learning system were guaranteed privacy and non-malicious handling of their information, they would be much more likely to participate. We analyze privacy dynamics in a learning system of buyers and a potentially price-discriminating seller, showing that an equilibrium arises which is privacy-protecting for the participants. Applying these insights to a more classic collaborative learning framework could spawn more widespread adoption of collaborative learning frameworks in practice. We study price-discrimination games between buyers and a seller where privacy arises endogenously—that is, utility maximization yields equilibrium strategies where privacy occurs naturally. In this game, buyers with a high valuation for a good have an incentive to keep their valuation private, lest the seller charge them a higher price. This yields an equilibrium where some buyers will send a signal that misrepresents their type with some probability; we refer to this as *buyer-induced privacy.* When the seller is able to publicly commit to providing a certain privacy level, we find that their equilibrium response is to commit to ignore buyers' signals with some positive probability; we refer to this as *seller-induced privacy.* We then turn our attention to a repeated interaction setting where the game parameters are unknown and the seller cannot credibly commit to a level of seller-induced privacy. In this setting, players must learn strategies based on information revealed in past rounds. We find that, even without commitment ability, seller-induced privacy arises as a result of reputation building, and we characterize the resulting seller-induced privacy and seller's utility under no-regret and no-policy-regret learning algorithms.

Future Directions. Integrating these insights on privacy into a collaborative learning framework, thus providing privacy guarantees for the participants, is a direction for future work.

While theoretically compelling, collaborative learning has yet to experience widespread adoption in practice. Designing systems which guarantee desired outcomes for participants is essential for incentivizing participation. This dissertation analyzes key areas of misalignment between system guarantees and client objectives and proposes solutions towards alignment. Much remains to be done.

Bibliography

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

[2] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, “Scaffold: Stochastic controlled averaging for federated learning,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 5132–5143.

[3] M. Even, L. Massoulie, and K. Scaman, “Sample optimality and *All-for-all* strategies in personalized federated and collaborative learning,” *arXiv preprint arXiv:2201.13097*, 2022.

[4] S. P. Karimireddy, L. He, and M. Jaggi, “Byzantine-robust learning on heterogeneous datasets via bucketing,” in *International Conference on Learning Representations*, 2022.

[5] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, “Taskonomy: Disentangling task transfer learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3712–3722.

[6] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, “Towards personalized federated learning,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[7] V. Kulkarni, M. Kulkarni, and A. Pant, “Survey of personalization techniques for federated learning,” in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, IEEE, 2020, pp. 794–797.

[8] F. Sattler, K.-R. Muller, and W. Samek, “Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints,” in *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32(8), 2021.

[9] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh, “Three approaches for personalization with applications to federated learning,” 2021.

[10] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, “An efficient framework for clustered federated learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 586–19 597, 2020.

[11] M. Even, L. Massoulié, and K. Scaman, "On sample optimality in personalized collaborative and federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 212–225, 2022.

[12] V. Smith, C.-K. Chiang, M. Sanjabi, and A. Talwalkar, "Federated multi-task learning," in *31st Conference on Neural Information Processing Systems*, 2017.

[13] T. Li, S. Hu, A. Beirami, and V. Smith, "Fair and robust federated learning through personalization," in *38th International Conference on Machine Learning*, 2021.

[14] L. Lamport, R. Shostak, and M. Pease, "The byzantine generals problem," *In Concurrency: the Works of Leslie Lamport.*, 2019.

[15] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] D. Yin, Y. Chen, R. Kannan, and P. Bartlett, "Byzantine-robust distributed learning: Towards optimal statistical rates," in *International Conference on Machine Learning (PMLR)*, 2018.

[17] G. Damaskinos, E. El-Mhamdi, R. Guerraoui, A. Guirguis, and S. Rouault, "Aggregator: Byzantine machine learning via robust gradient aggregation," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 81–106, 2018.

[18] R. Guerraoui and S. Rouault, "The hidden vulnerability of distributed learning in byzantium," in *International Conference on Machine Learning (PMLR)*, 2018, p. 35 213 530.

[19] K. Pillutla, S. Kakade, and Z. Harchaoui, "Robust aggregation for federated learning," in *IEEE Transactions on Signal Processing*, vol. 70, 2022, pp. 1142–1154.

[20] B. Baruch, M. Baruch, and Y. Goldberg, "A little is enough: Circumventing defenses for distributed learning," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] C. Xie, O. Koyejo, and I. Gupta, "Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation," in *Uncertainty in Artificial Intelligence (PMLR)*, 2020, pp. 261–270.

[22] S. P. Karimireddy, L. He, and M. Jaggi, "Learning from history for byzantine robust optimization," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 129, 2021, pp. 5311–5319.

[23] T. Wang, W. Cheng, D. Luo, W. Yu, J. Ni, L. Tong, H. Chen, and X. Zhang, "Personalized federated learning via heterogeneous modular networks," in *IEEE International Conference on Data Mining*, 2022, pp. 1197–1202.

[24] Y. Wu, S. Zhang, W. Yu, Y. Liu, Q. Gu, D. Zhou, H. Chen, and W. Cheng, "Personalized federated learning under mixture of distributions," in *Proceedings of the 40th International Conference on Machine Learning (PMLR)*, 2023.

[25] O. Marfoq, G. Neglia, L. Kameni, and R. Vidal, "Federated multi-task learning under a mixture of distributions," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 34, 2021.

[26] S. Vaswani, F. Bach, and M. Schmidt, "Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 89, 2019.

[27] Y. Arjevani, Y. Carmon, J. C. Duchi, D. J. Foster, N. Srebro, and B. Woodworth, "Lower bounds for non-convex stochastic optimization," in *Mathematical Programming*, vol. 199, 2023, pp. 165–214.

[28] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[29] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009. [Online]. Available: https://api.semanticscholar.org/CorpusID:18268744.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] K. Donahue and J. Kleinberg, "Optimality and stability in federated learning: A game-theoretic approach," *Advances in Neural Information Processing Systems*, vol. 34, pp. 1287–1298, 2021.

[34] K. Donahue and J. Kleinberg, "Model-sharing games: Analyzing federated learning under voluntary participation," in *35th AAAI Conference on Artificial Intelligence*, 2021.

[35] A. Blum, N. Haghtalab, R. L. Phillips, and H. Shao, "One for one, or all for all: Equilibria and optimality of collaboration in federated learning," in *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[36] X. Wu and H. Yu, "Mars-fl: Enabling competitors to collaborate in federated learning," 2022, pp. 1–11.

[37] S. P. Karimireddy, W. Guo, and M. Jordan, "Mechanisms that incentivize data sharing in federated learning," in *Federated Learning Conference at 36th Conference on Neural Information Processing Systems*, 2022.

[38] M. Werner, L. He, M. Jordan, M. Jaggi, and S. P. Karimireddy, "Provably personalized and robust federated learning," *Transactions on Machine Learning Research*, 2023.

[39] M. Han, K. Patel Kumar, H. Shao, and L. Wang, "On the effect of defections in federated learning and how to prevent them," 2023.

[40] J. Tirole, "The theory of industrial organization," vol. 1, no. 0262200716, 1988.

[41] A. Cournot, "Recherches sur les principes mathématiques de la théorie des richesses," 1838.

[42] C. Huang, S. Ke, and X. Liu, "Duopoly business competition in cross-silo federated learning," vol. 11, 2023.

[43] N. Tsoy and N. Konstantinov, "Strategic data sharing between competitors," in *37th Conference on Neural Information Processing Systems*, 2023.

[44] J. Bertrand, "Theorie mathematique de la richesse sociale," *Journal des Savants*, vol. 68, pp. 499–508, 1883.

[45] J. F. J. Nash, "The bargaining problem," *Econometrica*, vol. 18, pp. 155–162, 2 1950.

[46] H. F. v. Stackelberg, "Marktform und gleichgewicht," 1934.

[47] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," in *33rd Conference on Neural Information Processing Systems*, 2019.

[48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," 11, vol. 86, 1998, pp. 2278–2324.

[49] J. B. Foster and R. McChesney, "Surveillance capitalism," *Monthly review*, vol. 66, no. 3, pp. 1–31, 2014.

[50] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[51] K. Nissim, A. Bembenek, A. Wood, M. Bun, M. Gaboardi, U. Gasser, D. R. O'Brien, T. Steinke, and S. Vadhan, "Bridging the gap between computer science and legal approaches to privacy," *Harvard Journal of Law & Technology*, vol. 31, p. 687, 2017.

[52] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in Apple's implementation of differential privacy on macOS 10.12," *arXiv preprint arXiv:1709.02753*, 2017.

[53] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, IEEE, 2017, pp. 263–275.

[54] J. Dong, A. Roth, and W. J. Su, "Gaussian differential privacy," *arXiv preprint arXiv:1905.02383*, 2019.

[55] A. Solow-Niederman, "Information privacy and the inference economy," *Northwestern University Law Review*, vol. 117, p. 357, 2022.

[56] K. Ligett and K. Nissim, "We need to focus on how our data is used, not just how it is shared," *Communications of the ACM*, vol. 66, no. 9, pp. 32–34, 2023.

[57] A. Acquisti, C. Taylor, and L. Wagman, "The economics of privacy," *Journal of Economic Literature*, vol. 54, no. 2, pp. 442–492, 2016.

[58] A. Acquisti and H. Varian, "Conditioning prices on purchase history," *Marketing Science*, 2004.

[59] V. Conitzer, C. Taylor, and L. Wagman, "Hide and seek: Costly consumer privacy in a market with repeated purchases," *Marketing Science*, vol. 31, no. 2, pp. 277–292, 2012.

[60] R. Montes, W. Sand-Zantman, and T. Valletti, "The value of personal information in markets with endogenous privacy," *Center for Economic and International Studies*, vol. 13, no. 352, 8 2015.

[61] D. Fudenberg and J. M. Villas-Boas, "Behavior-based price discrimination and customer recognition," *Handbook on Economics and Information Systems*, vol. 1, pp. 377–436, 2006.

[62] J. Horner, "Reputation and competition," *American Economic Review*, vol. 92(3), pp. 644–663, 2002.

[63] C. Shapiro, "Premiums for high quality products are returns to reputation," *Quarterly Journal of Economics*, vol. 98(4), pp. 659–679, 1983.

[64] J. Ely and J. Valimaki, "Bad reputation," *The Quarterly Journal of Economics*, vol. 118(3), pp. 785–814, 2003.

[65] O. D. Hart and J. Tirole, "Contract renegotiation and coasian dynamics," *The Review of Economic Studies*, vol. 55, no. 4, pp. 509–540, 1988.

[66] S. Ichihashi, "Online privacy and information disclosure by consumers," *American Economic Review*, vol. 110, no. 2, pp. 569–595, 2020.

[67] M. K. Camara, J. D. Hartline, and A. Johnsen, "Mechanisms for a no-regret agent: Beyond the common prior," *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, 2020.

[68] Y. Deng, J. Schneider, and B. Sivan, "Strategizing against no-regret learners," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[69] N. Haghtalab, C. Podimata, and K. Yang, "Calibrated Stackelberg games: Learning optimal commitments against calibrated agents," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[70] D. P. Foster and R. V. Vohra, "Calibrated learning and correlated equilibrium," *Games and Economic Behavior*, vol. 21, no. 1-2, p. 40, 1997.

[71] R. Arora, M. Dinitz, T. V. Marinov, and M. Mohri, "Policy regret in repeated games," *Advances in Neural Information Processing Systems*, 2020.

[72] T. Lattimore and C. Szepesvari, *Bandit Algorithms*. Cambridge University Press, 2020.